WeRateDogs Twitter Archive - Act Report

> **Stephanie Anderton**
>
> DAND Project 8
> *December 2018*

WeRateDogs Data

The **WeRateDogs** *Enhanced* Twitter archive contains data extracted from 2356 of the 5000+ tweets from the @dog_rates twitter account, posted between November 15, 2015 and August 1, 217. This data is comprised of dog ratings that were taken from the text of the tweet along with the dog name and dog stage if present.

The retweet count and favourite count for each tweet were not included in the enhanced archive, and so I had to download this additional data from the twitter account using the tweet ID from the archive.

Along with the Twitter data, I also downloaded an image predictions file from Udacity servers containing the top 3 predictions for dog breeds based on the images from the tweets (there can be up to 4).

Wrangling Data

Before I could begin the analysis, the data had to be wrangled into shape to make it easier. I assessed the data both visually and programmatically for quality and tidiness; the quality of data is determined mainly by looking at several aspects or dimensions to ensure that it is complete, valid, accurate and consistent.

After cleaning many of the issues found during the assessment, there were about 1950 tweets with good quality data.

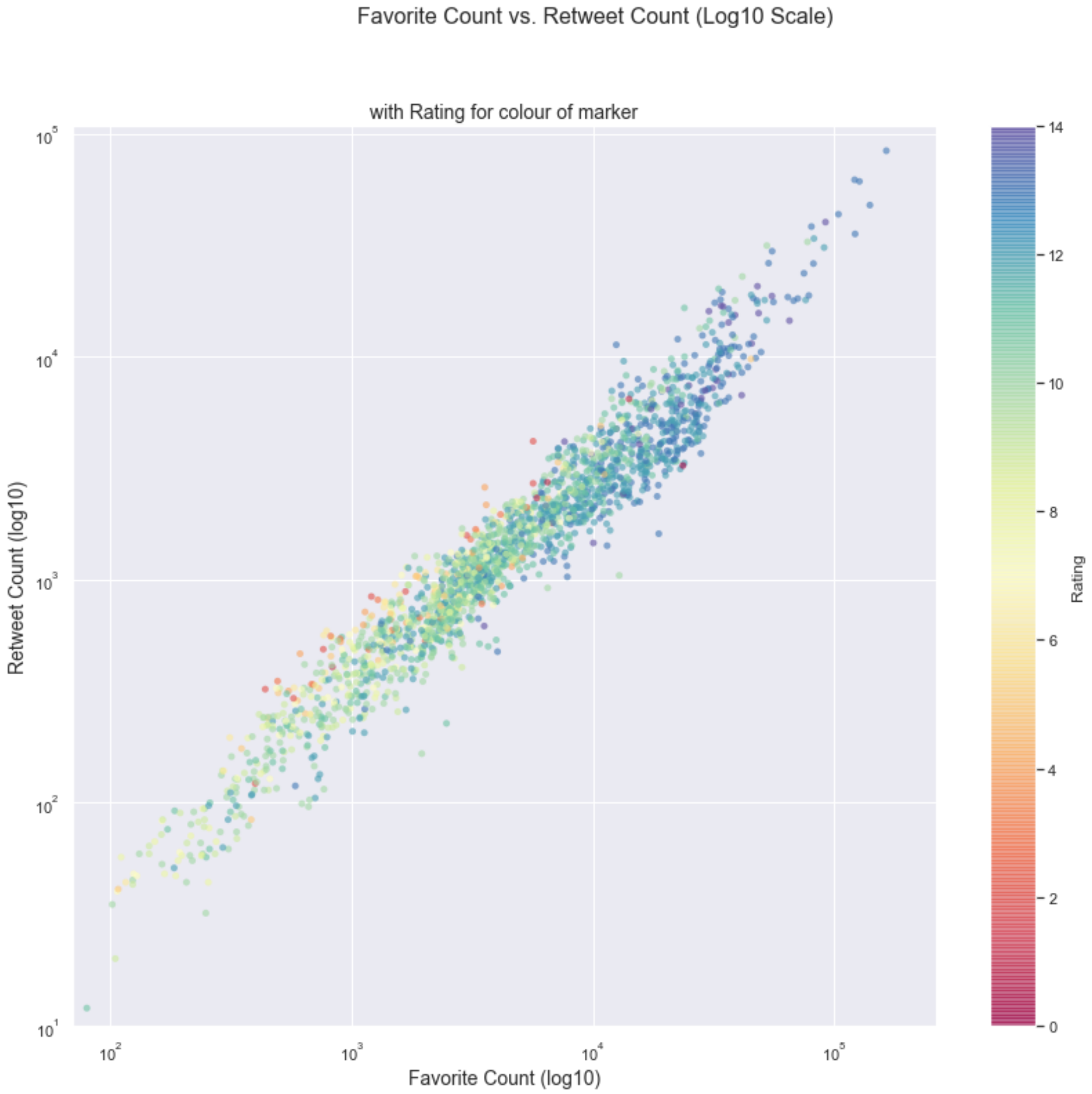
Insights

Favourite Counts Are Higher Than Retweet Counts

One of the simplest insights was that ALL tweets have higher favourite counts than retweet counts. From personal experience, I am not really surprised; I am more likely to favourite a tweet than retweet it.

Strong Relationship Between Favourite Count And Retweet Count

There is a strong relationship between a tweet's retweet counts and favourite counts; in other words if a tweet has a high favourite count it will also have a high retweet count. Plotting these two values against each other produced the plot below, and clearly shows a high correlation by the very tight collection of points along the angled, straight line.



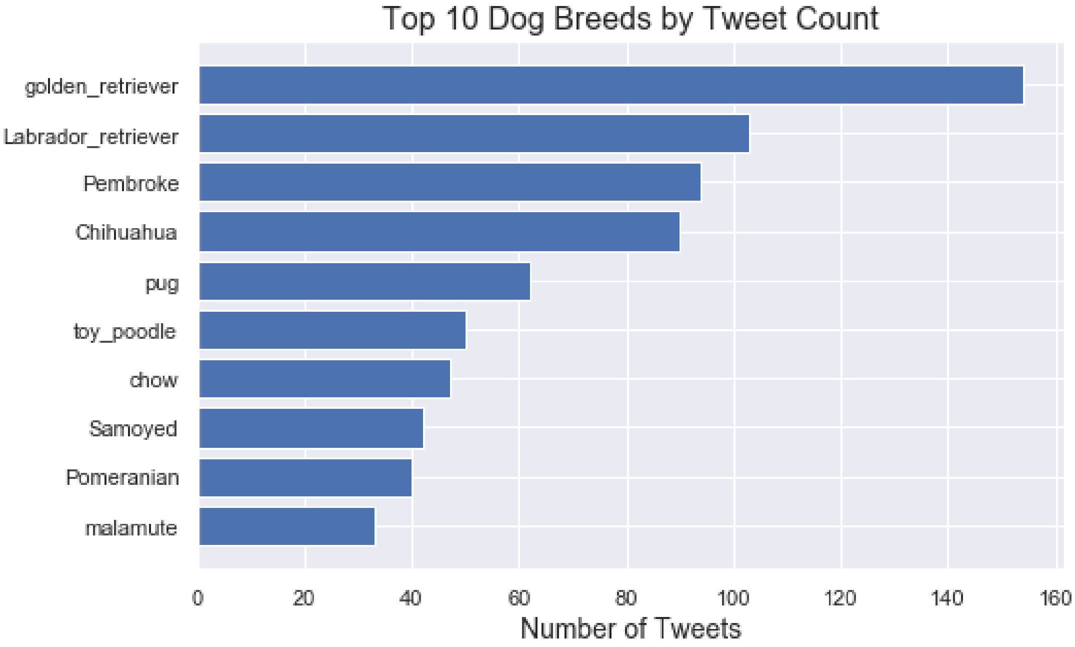
Higher Ratings Mean Higher Retweets and Favourites

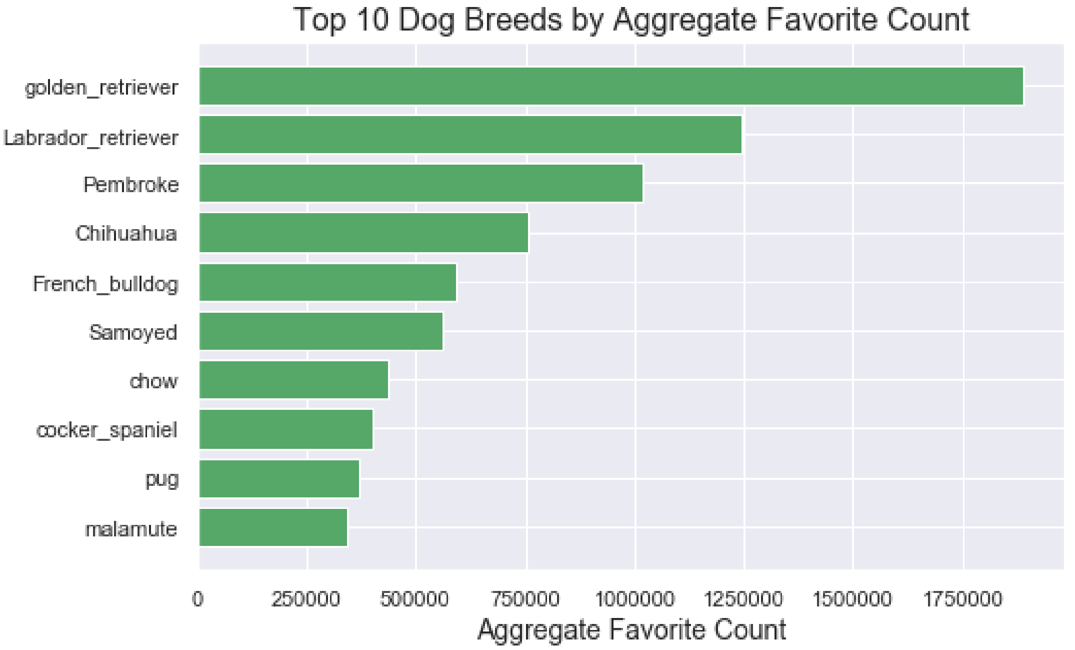
Also, when I added *colour* to each tweet (dot) to denote the rating (blue is highest, red lowest), I noticed that there are many more blue dots in the upper right, where favourite counts and retweet counts are highest. Conversely, there are more mid-level ratings (yellows, oranges and greens) in the lower left where the retweet and favourite counts are lower. This does make sense as a tweet would be retweeted and/or favourited if the dog (the subject) is remarkable in some way, garnering a rating to reflect this.

Folks Love Retrievers

When the images from the tweets were fed through a neural network to predict the dog breed, we got predictions for 85% of them, and it identified 113 different breeds. The confidence of the predictions also ranges a lot, and many tweets do have low levels. However, only 300 or so tweets' images didn't look like 'dog'; I wonder what those images were about!

Grouping all the tweets by the dog breed showed that the top 4 breeds of dogs with the *most number of tweets* are also the same as those with the *highest total favourite count* (see the plots below), and the top 4 are ranked in the same order. There are 8 breeds in common: golden_retriever, Labrador_retriever, Pembroke, Chihuahua, pug, chow, Samoyed, and malamute.

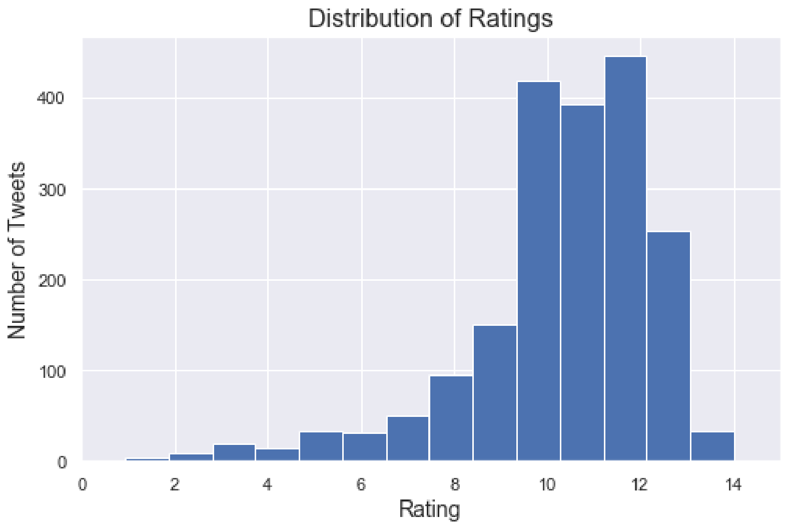




Folks sure love retrievers!

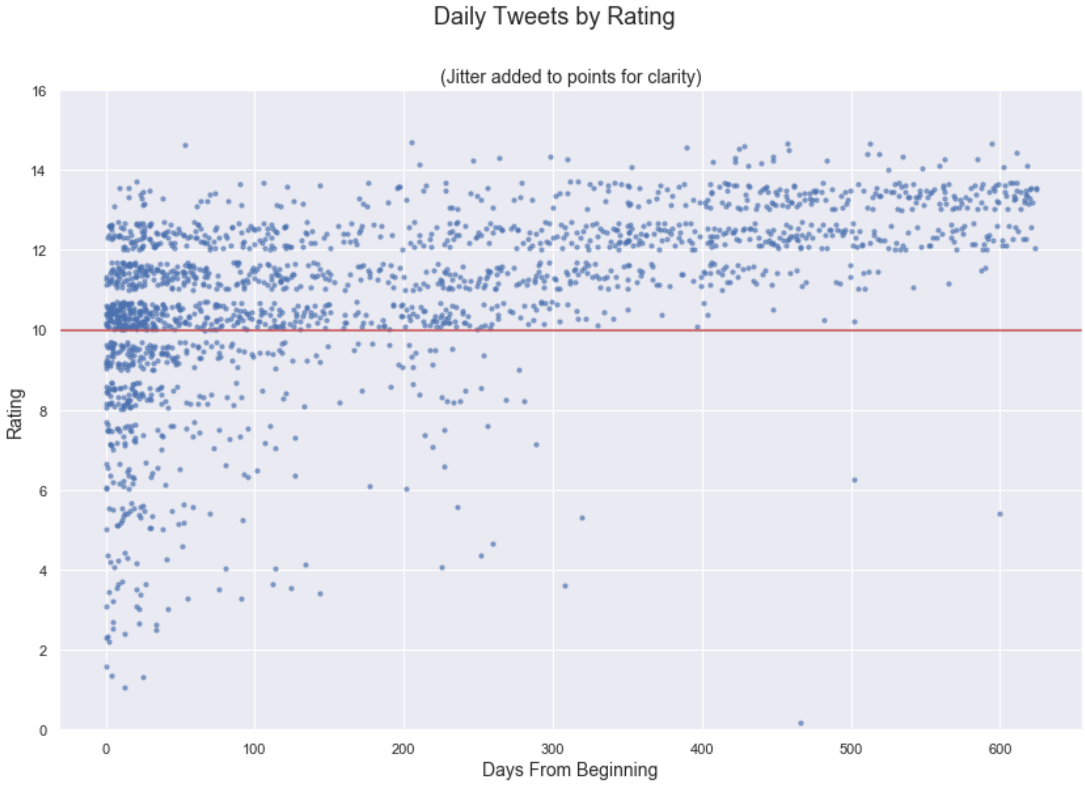
So Many High Ratings

The distribution here shows just how many tweets there are with really high ratings; in fact there are over 1500 tweets with ratings of 10 and up. That’s 79% of all tweets!



They're Good Dogs Brent!

After the account had been running for about 300 days almost all the tweets were rated 10 and up (above the red line). Practically every tweet!



As the twitter account replied to one user, "They're good dogs (Brent)!"