Insights from Data Wrangling the WeRateDog twitter account

Premise: all the insights and visualizations in this analysis has been obtained from data gathered from three sources: a twitter_archive.csv file with data about WeRateDogs tweets, a file gathered from an url provided by Udacity with image predictions about the content of images in tweets from WeRateDogs and finally from Twitter with the help of the Tweepy API. The data has been assessed and cleaned creating a merged dataframe of all the sources.

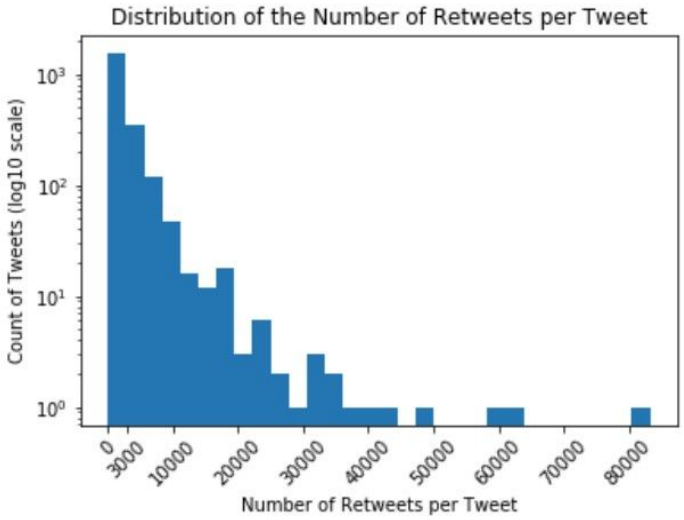
WeRateDogs (https://twitter.com/dog_rates), is a well known twitter account where images of dogs are posted usually with humorous comments and a score given to the dogs.
The scores are ideally expressed in tenths, but many dog owners love their pet so much that very often the score is greater than 10/10, as in 14/10: this is commonly accepted on WeRateDogs.
Not every image posted on WeRateDogs is really a dog: the idea behind this is that if the dog is the man's best friend, many other animals or sometimes objects (like a fan in summer) can be somewhat a sort of "dog", so "dog" is a positive word than can mark more than simply dogs.
Sometimes dogs and "not dogs" are called with sort of surnames like "doggo" (adult dog), pupper (young not yet adult dog), puppo (very young dog) or floofer (a flurry dog).

The success of tweets is often determined by the number retweets and favorites they receives, let's look at retweets and favorites on WeRateDogs.

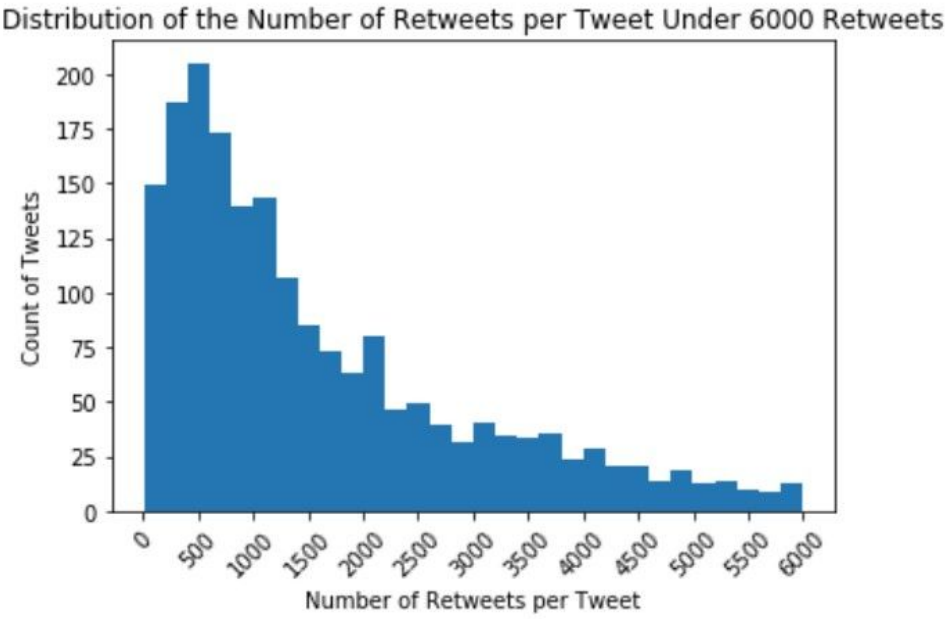
The first insight about WeRateDogs is that the number of retweets is extremely variable, but most of tweets get less than 3000 retweets and puppo dogs are the most retweeted.

Ranging from a minimum of 11 to a maximum of 83275, no tweet remain totally unnoticed on WeRateDogs, but there is sometimes a huge difference in popularity, measured as number of retweets, between tweets: the average number of retweets is about 2693, but the standard deviation is very high, about 4710.

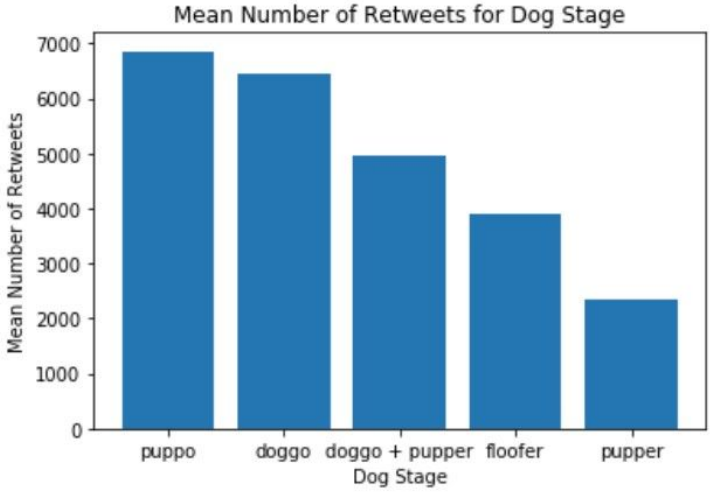
Below, plot of the distribution of retweets with the number of tweets on a log10 scale.



Below instead a more close look at the distribution of retweets for tweets under 6000 retweets, including so about 90% of all the tweets.

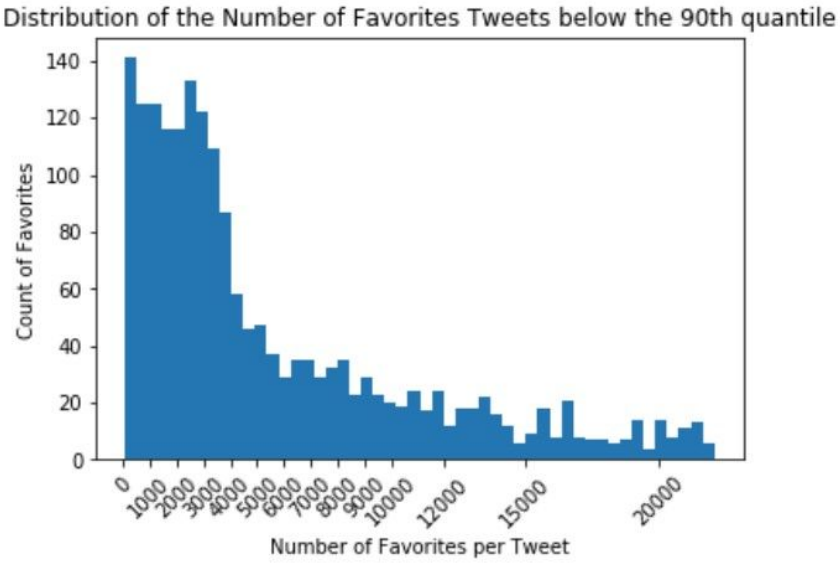


Looking at retweets for dog stage, puppo dogs are the most retweeted, as shown in the plot below:

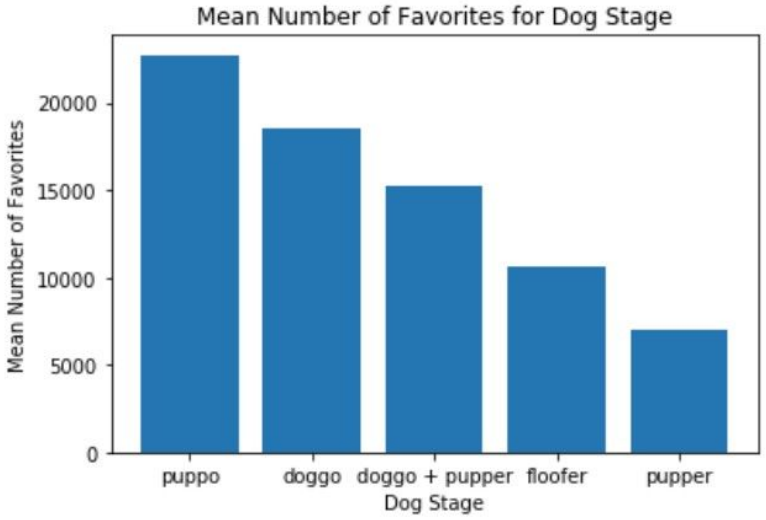


Second insight : the number of favorites per tweet is on average higher than the number of retweets, but it is again extremely variable. Puppo dogs are again favorite stars on WeRateDogs.

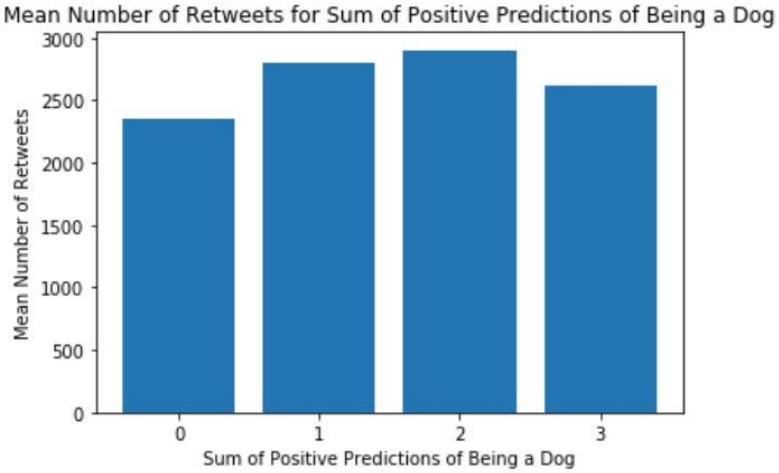
The plot below shows the distribution of favorites for 90% of the tweets on WeRateDogs, excluding so the tweets with a very high number or retweets.



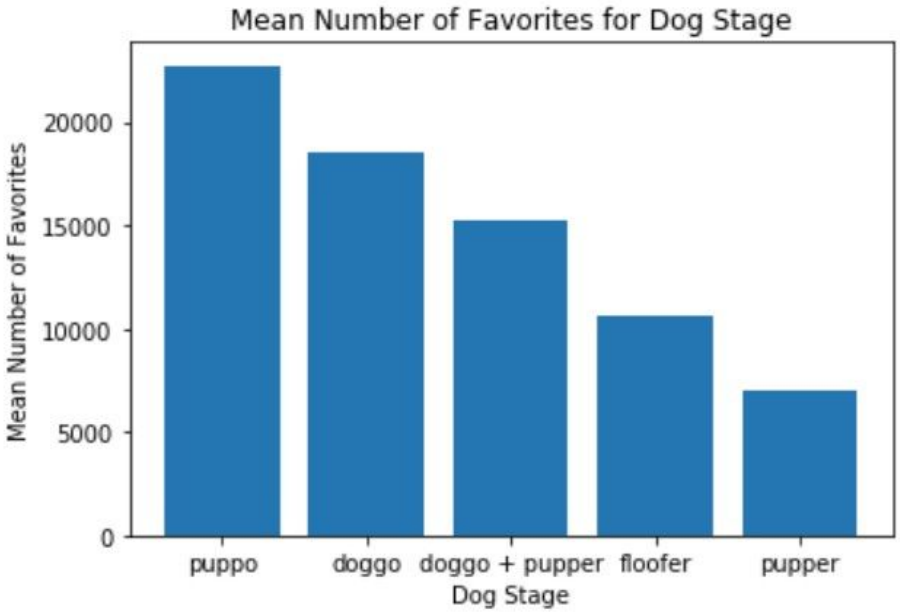
On average the number of favorites is higher than the number of favorites, as expected on Twitter, ranging from a minimum of 78 to a maximum of 163767, with an average of about 8728 favorites per tweet and a standard deviation of about 12669 (so very high). Puppo dogs seem to be the most favorites, as they were the most retweeted, as shown below:



Udacity has feeded a neural network with images from WeRateDogs: the neural network has given three possible predictions about the contents in every image, giving for every of them the probability of being a dog. So every image can receive up to three possible classifications as dogs. I calculated the sum of the number of times an images has been classified as dog (so ranging from 0 to 3), and below you can see the number of retweets for every sum: being zero for the images with less probability of being a dog and 3 for the ones with the highest probability of being a dog.



Below a plot similar to the previous one, but for the number of favorites.



Insight 3: The most probable dog isn't always the most probable favorite or the most retweeted, but pictures that are probably something different from dogs are the less ones

Not every image posted on WeRateDogs is really of a dog: sometimes it is of other animals, but it can also be of whatever nature (even a fan), but they get scores, retweets and rates nonetheless.
In theory pictures classified as almost certainly as dogs (three classifications as dog) should have greater success on average than pictures of only partially classified as dogs (2 predictions of being a dog on 3 in total), getting more retweets and favorites, but instead the tweets with the average greatest success are the one with a total number of predictions of being a dog of 2 (even if they are very similar in average retweets to the ones with a 3 classifications as dog, 2902 versus 2804 retweets and 9133 versus 9045 favorites).
Sometimes pictures contain not only a dog, but something else that is very visible near the dog and that is classified by the neural network as the possible main content of the picture: maybe that something else gives the dog the right "light" to be noticed by twitter users (this is only a theory!) in pictures that get only 2 predictions of being a dog.
As expected, picture that are most probably not a dog, with zero predictions of being a dog, get the less retweets and favorites of all on average.

Insight 4: the scores given to dogs aren't correlated to the number of retweets and favourites

No matter how much good or bad are the scores given to the dogs, the number of retweets and favorites aren't correlated to the scores: the pearson correlation coefficient between number of retweets and score is 0.016 and between favorites and score it is 0.015, so very near to zero.
It seems brave and good dogs need to earn their success on their own!