# WeRateDogs Data Analysis

**Author: Federico Finetti**
**Date: 10 February 2019**

This analysis is about the popular Twitter account WeRateDogs, where there are pictures of dogs (in theory), with relative ratings.
The particularity of this rating system is that very frequently appear values with the numerator larger than the denominator. Also, the denominator isn't constant, so the different dogs aren't so easy to compare sometimes.
I achieved consistency by using the result of the division (numerator / denominator) as the ultimate rating value, so that all the dogs can now be compared easily.

Once the Wrangling phase was concluded, and I could analyse datasets properly assessed and cleaned from Quality and Tidiness issues, was time for the real data analysis.

A quick look to the dataframe showed that the average dog rating was 1.224, with a standard deviation of 5. The average rating greater then one, confirms that, as the motto of the website is, ***they are good dogs, Brent!*** …but why such a big standard deviation?

First of all, it's very important to investigate the outliers, to understand the reasons why they have those values and to understand if they are appropriate for the analysis and the model we are building.

Atticus, here on the right has a rating of 177.
No questions that it's a really lovely dog, but such a high rating is probably due to the irrationality due to the special patriotism in a special day in the US (photo published on the 4t of July, which is the Independence Day in USA), therefore I believe this value would be a bit misleading for the analysis.

Here on the left you can see another very popular Dog(g), with rating 42.
No doubt he is a great artist but again I think it's not appropriate for the analysis, therefore this datapoint was removed.

It's also interesting to investigate the dogs that had the lowest ratings, to understand the reasons, if any. I discovered that some records have a very low ratings because they aren't actual dogs.
In particular, three "dogs" having rating 0.1 (lowest value of the dataset) were:

1) A fan on a sofa.

2) A really lovely dog because, as it's clear from the comment "*What kind of person sends in a picture without a dog in it? 1/10 just because that's a nice table*", it wasn't seen as it was perfectly camouflaged in the carpet.

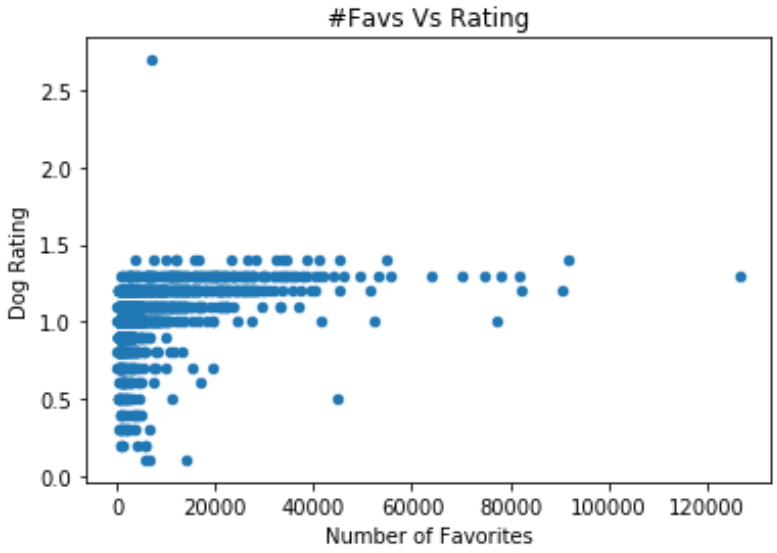

3) Again, not a dog but an actual chicken.

After having removed the top three outliers, with values over 3.42, the average dog rating decreased to 1.05 approximately, but the standard deviation literally dropped at 0.22.

If you are a bit in a rush and just want to see the very best dogs (the top 16%, which are the ones with a rating of one standard deviation above the average or more), you should look at dogs with ratings of at least 1.27.

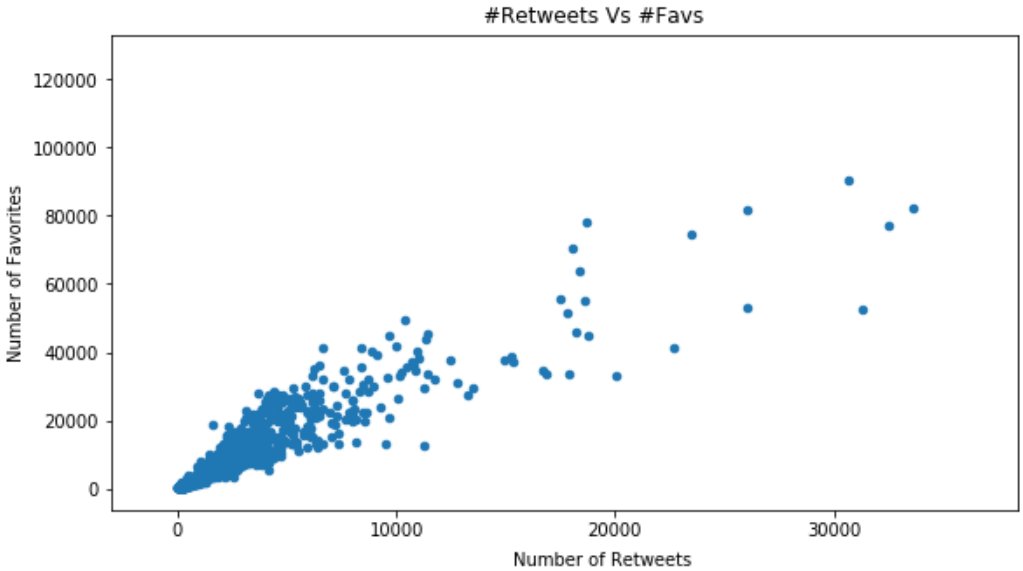
With regard to the Cause-Effect analysis, let's think and formulate a theory.

I thought it would have made sense to assume that the rating is positively correlated with the number of "favorites".
So I calculated the correlation between these two dimensions and, surprisingly, there was only a weak positive correlation (0.38) between them.



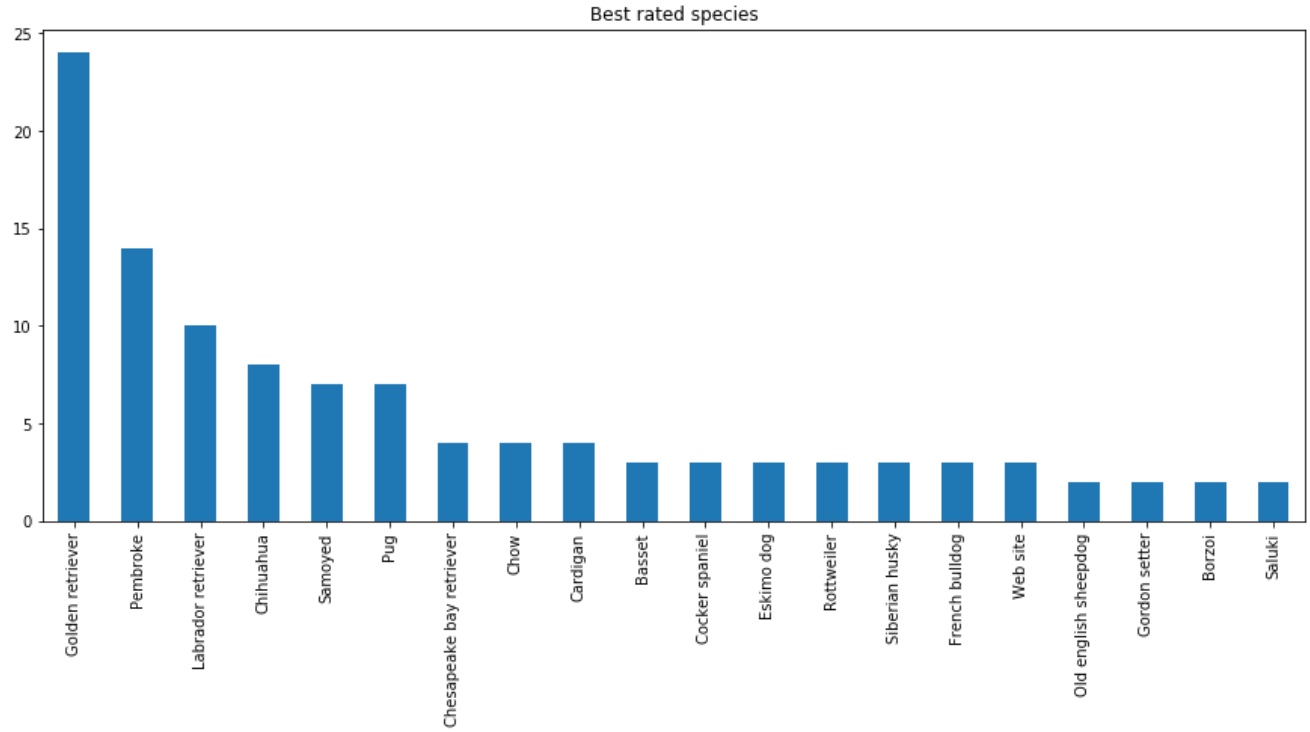

The reason for this might be that a dog not so beautiful attracts funny comments and are those comments that attract more favorites.

This theory seemed to find some statistic support when we looked at the correlation between the number of retweets and the number of favorites, which was strongly positive (0.93).
This is probably due to the fact that the more are the retweets, the more likely is that some of them are really funny or interesting, therefore attracting favorites.



I wanted then to test, with a Multiple Linear Regression, if the categorical variable "stage" impacted the rating.
An R-squared value of 0.003 indicated that only an extremely small part of the dog rating value was explained by the dog stage. The p-values of the stages higher then the pre-set alpha-level at 0.05 confirmed that we had enough confidence to affirm that the stage had no impact on the rating.

Finally, I wanted to analyse which were the breeds of dog with the highest rating.
So I analysed the very best dogs in the top 16% rating, looking at how many they were for each breed.




We found out that the Golden retriever is by far the most appreciated type of dog, with Pembroke in the second position and Labrador retriever on the lowest step of the podium.

In particular, the Golden retriever average rating was 1.154, versus a population average rating of 1.05.
Adopting one of those dogs would probably increase the likelihood of getting a rating higher than average.

…no surprise looking at lovely Ollie here on the right…