Udacity: Data Analyst Nanodegree

Project: Wrangle and Analyse Data

By: Laszlo Rado 03-08-2020

Analysis Report

Context of the project:

WeRateDogs is Twitter account with 8.8M followers and according to their profiles they are:

"Your Only Source For Professional Dog Ratings Instagram and Facebook"

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Our Goal:

Gather, assess and clean data from multiple sources to create interesting and trustworthy analyses and visualizations about WeRateDogs Twitter account!

For further details about the data wrangling process please see the wrangle_report.pdf

Topics to Analyse:

I thought it would be interesting to look at what are the most popular dog breeds based on different measures like the proportion of tweets related to that specific breed or the proportion of favourites and re-tweets a specific breed received?

After we have figured out what are the most popular breeds we can do some further analysis! Let's figure out whether one type of popularity implies other type of popularity as well!

What do I mean by that?

We will find out if a more popular dog, based on the proportion of tweets are also more likely to get more favourites and re-tweets! We are going to create some interesting visualisations and linear regression models to figure it all out!

And if we are already there, we will see whether we can predict the number of re-tweets based on the number of favourites!

So what do you think about it? Do you think more popular dog breeds are more likely to be favourited? Do you think there are any correlation between the number of likes and the number of re-tweets?

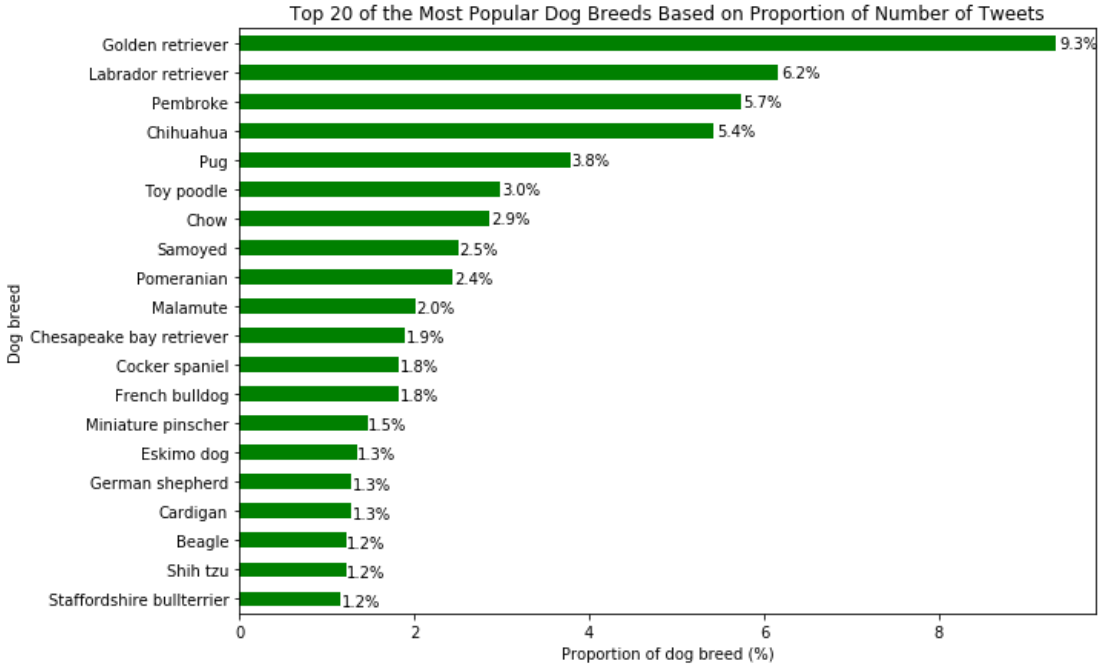
And most importantly! What is your initial prediction about the most popular dog breed?!

Turn to the next page to figure that out! Let's see those puppies!

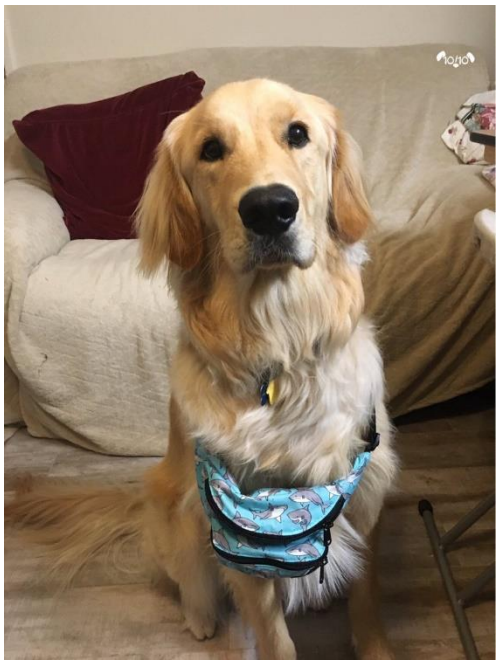
A heavily favourited puppo.

Top Lists

I have created a visualization of the 20 most popular dog breeds by the number of tweets that specific breed has to get a feeling about which breeds are on the top.



And guess what! The most popular dog breed is… drumbeat!



Obviously the Golden retriever! This didn't came as a surprise did it?

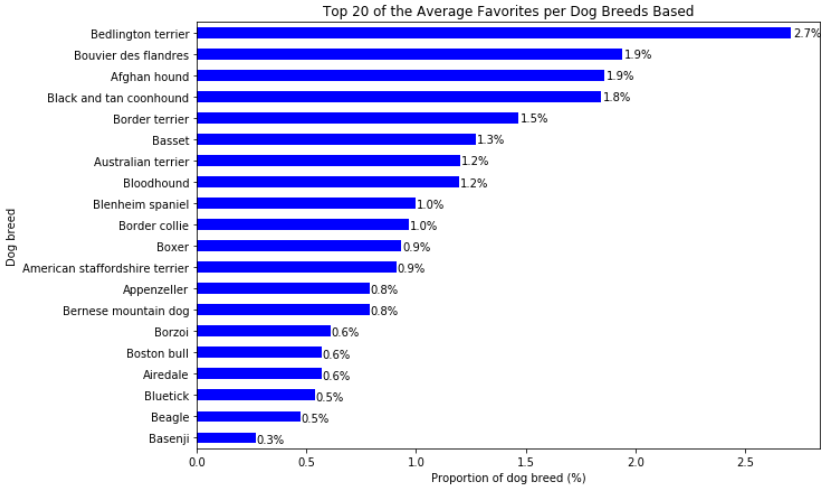
Together with their close friends the Labrador retrievers, they occupy the first two places on the list and responsible for over 15% of the tweets. This is a huge number out of all the dog breeds out there!

The Golden retriever is almost 3 times as more popular than the Pug, the 5th place on the list! This is not even close! The Golden retrievers easily win the first test. It looks like these dudes are really popular!

But are they similarly popular when it comes to favourites and re-tweets?

Let's figure that out!

1) Now let's look at what proportion of the favourites and re-tweets these dogs receive to see it whether they are also the most popular when comes to liking them?



We can clearly see that the average favourites per breed is a completely different list than the popularity based on the number of posts per breed. Our top two retrievers are not even in the list!!!

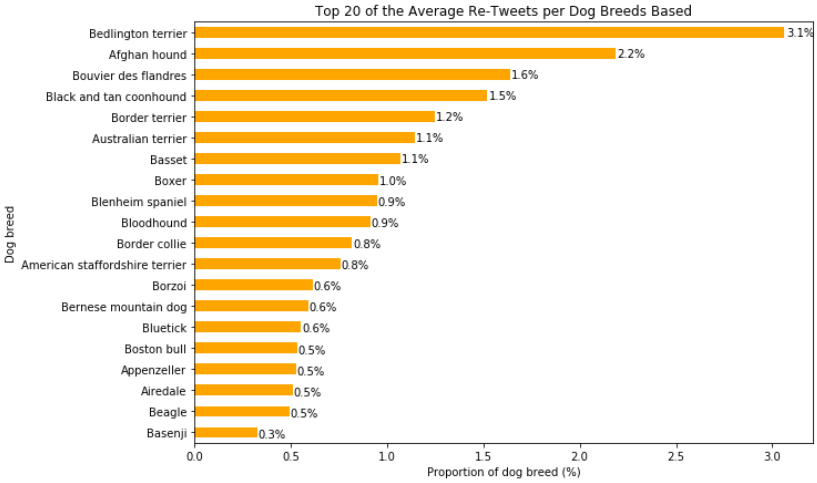
Our winner here is the Bedlington terrier!

(Okay, this is not so much a Bedlington terrier, but our Neural Net believed it is. So let's just don't hurt her feelings and leave it here!)

It is also clear that the average favourites per breed is much more evenly distributed than the popularity just by looking at the range of values (we only see the maximum in the above lists, but the minimum can't be smaller than 0 so we have a sense about the differences in the two range).

So the Bedlington terrier is a much tighter winner of this club, than the Golden retriever of the previous one.

We will confirm this later by calculating the standard deviation for these variables.



Let's now look at the distribution of top 20 re-tweets.

It resembles the chart for the average favourites per breed very much, which already suggests that we can expect some correlation between the favourites and the retweets, but not so much between the proportion of posts.

Standard Deviations

If we look at the standard deviations of the three variables we have analysed so far, we can indeed see that the proportions of average favourites and re-tweets are much more similar to each other than to the proportions of posts per breed. They have a very similar standard deviation.

They are also much more evenly distributed than the proportions of posts per breed. The different values vary closer around the mean. We can see this from the smaller standard de

viations.

So, when it comes to likes and re-tweets the dog breeds are not that different than when it comes to the proportion of posts!

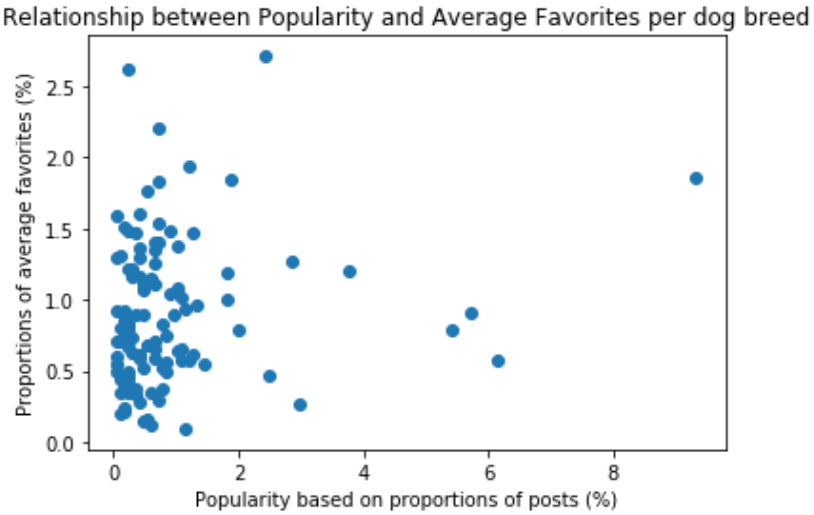
Measure	Standard Deviation
Proportion of posts per breed	1.320233
Proportion of average favorites per breed	0.512041
Proportion of average re-tweets per breed	0.531482

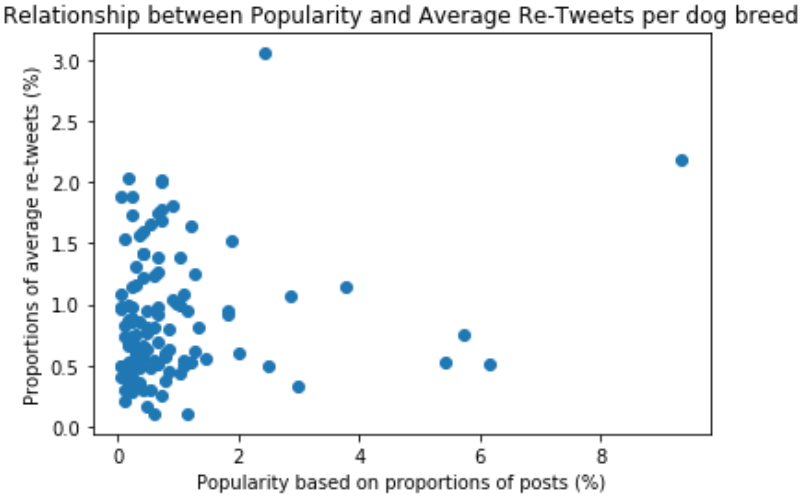
And to not leave any page without a dog!

This little dude is a Vizsla. A breed originated from my country, Hungary.

Correlations between popularity and re-tweets / favouries

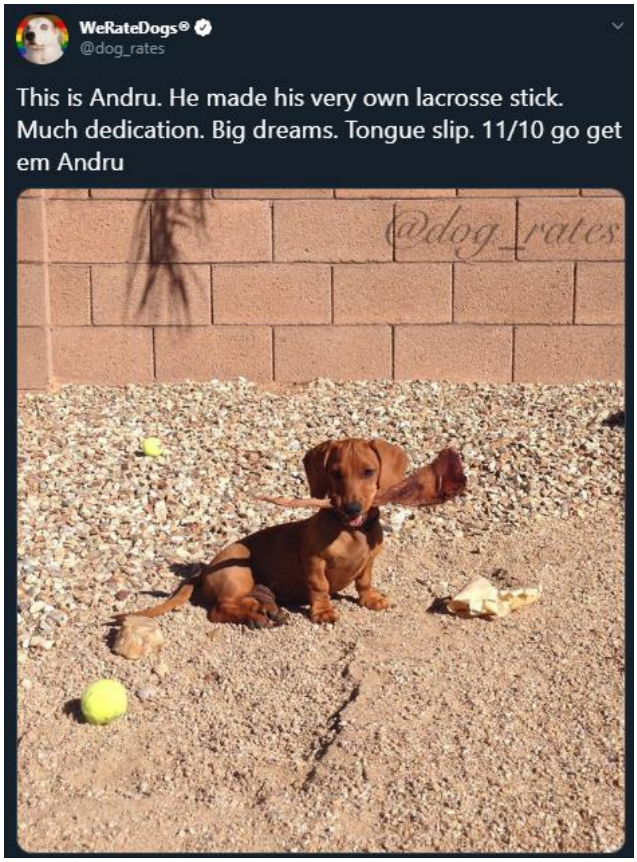
So let's now visualize the relationship between the Popularity of a dog breed based on the proportion of tweets it has, against the proportion of number of likes and re-tweets that specific dog breed has received.





We can clearly see from the above scatter plots that there is no correlation between the average re-tweets or the average favourites and the popularity of the dog breed.

This means that a more popular dog breed is not more likely to be re-tweeted or favorited.



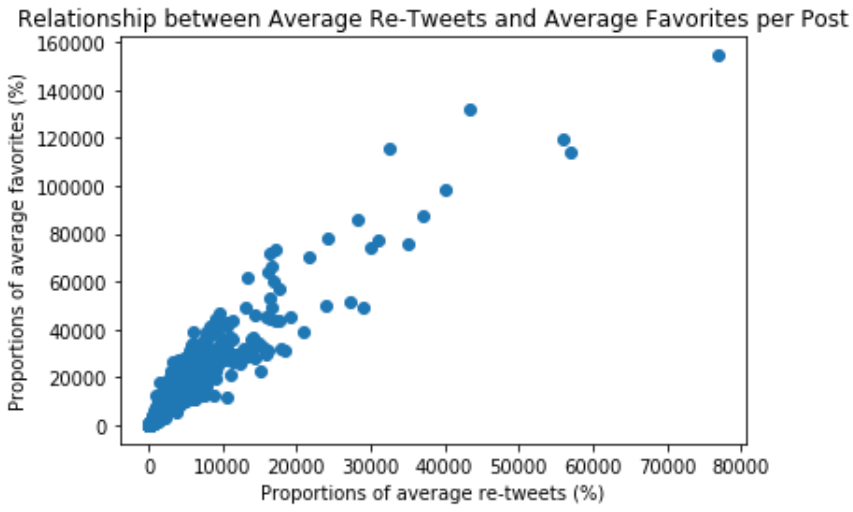
Calculating a linear regression model between the Popularity of a dog breed based on the number of tweets it has against the number of favourites the same dog breed receive on average we can confirm our visual observation.

The linear regression model returned an R-squared value of 0.027 which means the variance in the average likes of dog breed only explains 2.7% of the variance in the popularity of the popularity of the dog breed based on the number of posts. This implies very low, practically no correlation between the two variables.

We can also see that the p value for our coefficient is 0.081 (which is higher than 0.05), meaning the independent variable is not statistically significant in relating to the dependent variable.

This confirms our visual observation, that the popularity of a dog breed based on the number of posts does not imply higher amount of likes for the same dog breed.

Correlations between tweets and favourites



Let's look at whether there is a correlation between the number of re-tweets and the number of favourites!

Based on our analysis so far, we already expected that there should be correlation, and indeed, our visualisation on the left confirms it very well. We can see a strong positive correlation here.

Let’s confirm our conclusion based on the scatter plot by looking at the results of the regression model I created between the favourites and re-tweets!

We can see a high R-squared value of 0.883 which means the variance in the average favorites explains 86.3% of the variance in the re-tweets of a posts. This is implies very strong positive correlation between the two variables.

We can also see that the p value for the coefficient is 0.000 (which is lower than 0.05), meaning the dependent variable is statistically significant in relating to the independent variable on 95% confidence level.

This confirms our visual observation, that more favorites imply more re-tweets.

Our final conclusions and insights:

1) The most popular dog breed is Golden Retriever, but interestingly this breed is not in the lists of top20 most favorited and top20 most retweeted breeds! The top liked and re-tweeted breed is the Bedlington terrier.

2) The popularity of a dog breed based on the proportion of posts that specific dog breed has does not imply more favorites and re-tweets (one kind of popularity (number of posts) does not imply other type of popularity (favorites and re-tweets).

3) More favorites imply more re-tweets. Based on the favorite_count coefficient (0.3343) we can expect that the number of re-tweets of the average post is 33.43% of the favorites received.

4) Dogs are awesome! Thank you for reviewing my project! :)