Project 2: Wrangle and Analyze Data

Data Analysis Report

Prepared By: Suzan Hamza

Introduction

This report briefly communicates the insights and displays the visualization(s) produced from wrangling *WeRateDogs* Twitter data.

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 and numerators, though, are almost always greater than 10!

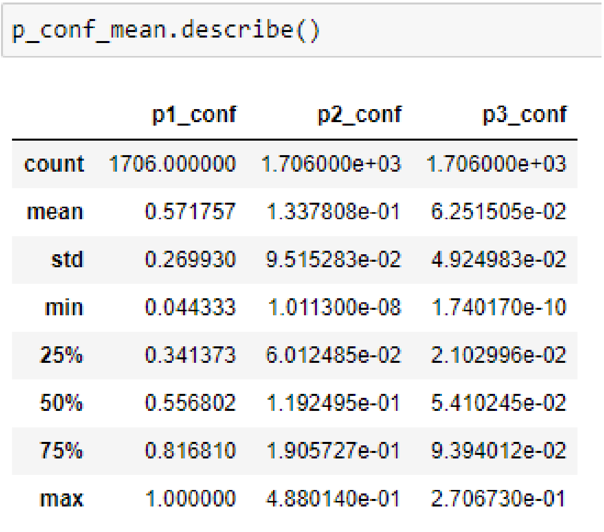
After wrangling the data in the first part, let's dive deeper into the cleaned dataset to discover some interesting insights about dogs.

Q1: What are the most confident image predictions (p1, p2 or p3)?

The WeRateDogs Twitter archive was run through a **neural network** that can classify breeds of dogs, and the algorithm produced three different predictions (`p1`, `p2` and `p3`).

We would like to know which one of these predictions had the highest confidence level, to use the most confident prediction in our analysis.

After displaying the basic statistical details for the confidence level of each prediction:



```
p_conf_mean.describe()
```

	p1_conf	p2_conf	p3_conf
count	1706.000000	1.706000e+03	1.706000e+03
mean	0.571757	1.337808e-01	6.251505e-02
std	0.269930	9.515283e-02	4.924983e-02
min	0.044333	1.011300e-08	1.740170e-10
25%	0.341373	6.012485e-02	2.102996e-02
50%	0.556802	1.192495e-01	5.410245e-02
75%	0.816810	1.905727e-01	9.394012e-02
max	1.000000	4.880140e-01	2.706730e-01

Figure 1 Basic statistical details for the confidence level of each prediction

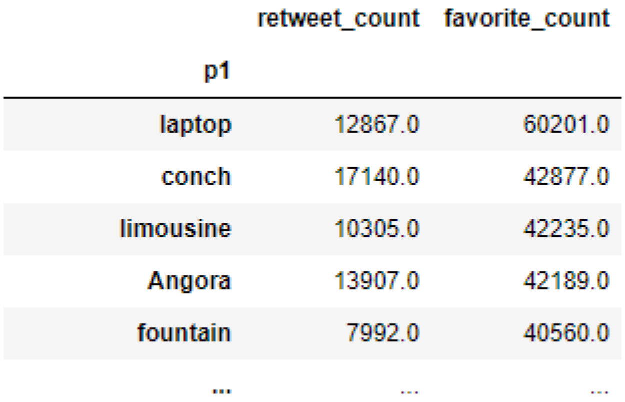
We can see that the first prediction `p1` tends to have the **highest average confidence** level (**57%**), with a **minimum** of (**4%**) and **maximum** of (**100%**).

Hence, we will use the first prediction `p1` for the rest of our analysis.

Q2: What are the most popular dog breeds?

The retweet and favorite counts for each dog breed will determine the most popular dog breeds that attract the largest number of Twitter followers.

However, after grouping the average retweet and favorite counts for the first dog breed prediction `p1`, it turns out that the highest average retweet and favorite counts were to objects (like *Laptop*, *conch*, *Limousine*, etc.) and not for dogs!

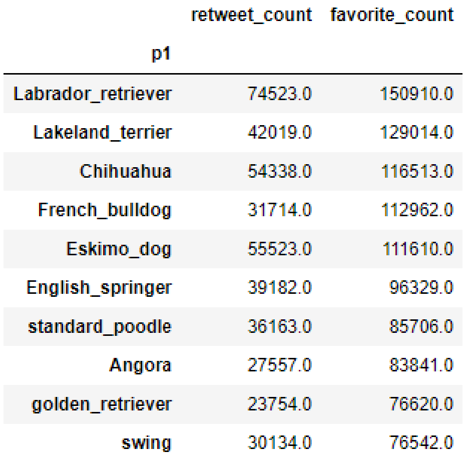


	retweet_count	favorite_count
p1		
laptop	12867.0	60201.0
conch	17140.0	42877.0
limousine	10305.0	42235.0
Angora	13907.0	42189.0
fountain	7992.0	40560.0
...	...	...

Figure 2 Highest average retweet and favorite counts

This indicates that not all image predictions were correct, and we will investigate that further in another section.

In the meantime, we will use another statistic `` `maximum` `` to discover our most popular dog breeds.



	retweet_count	favorite_count
p1		
Labrador_retriever	74523.0	150910.0
Lakeland_terrier	42019.0	129014.0
Chihuahua	54338.0	116513.0
French_bulldog	31714.0	112962.0
Eskimo_dog	55523.0	111610.0
English_springer	39182.0	96329.0
standard_poodle	36163.0	85706.0
Angora	27557.0	83841.0
golden_retriever	23754.0	76620.0
swing	30134.0	76542.0

Figure 3 Top Ten Dog Breeds

Hurrah! Most of the breeds this time are for dogs. The following bar plot visualization shows that the most popular dog breeds, are *Labrador retriever*, *Lakeland terrier*, *Chihuahua*, etc.

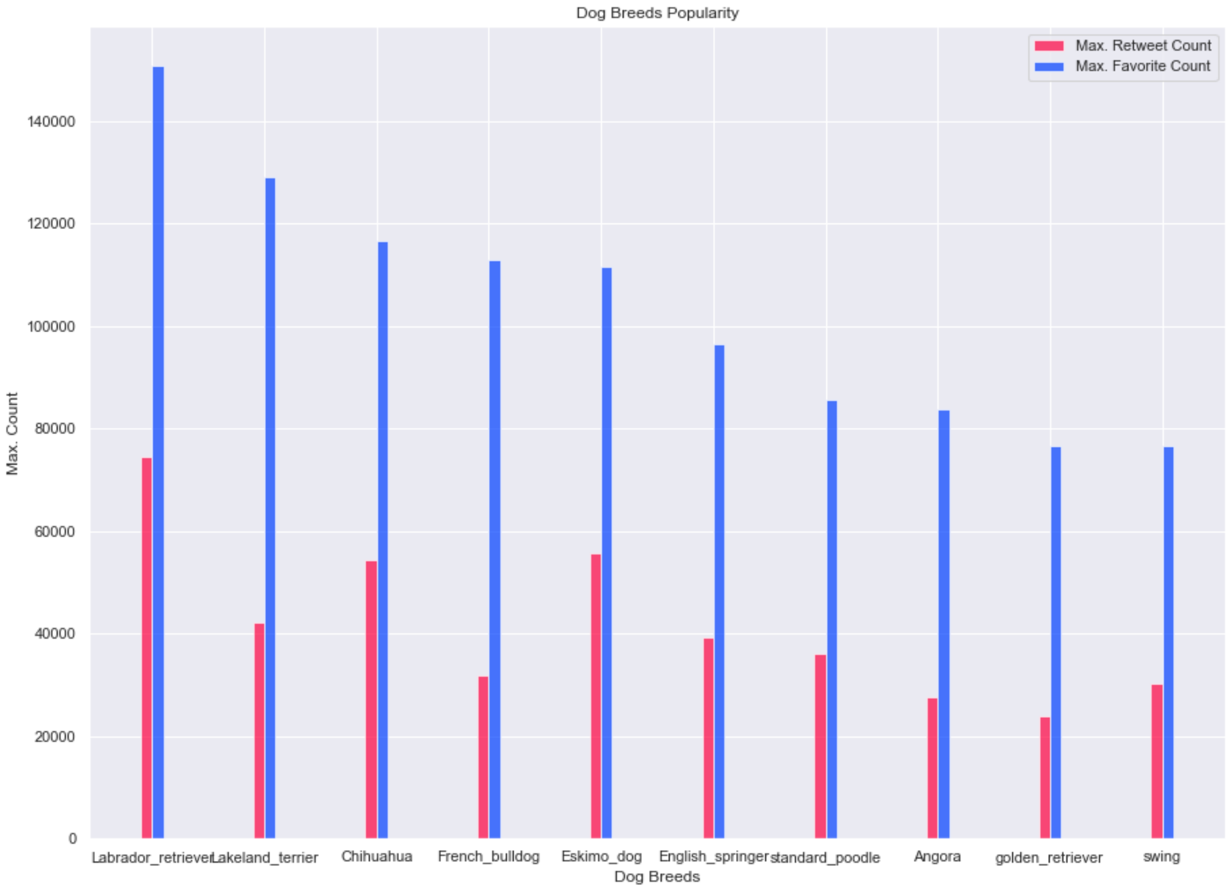
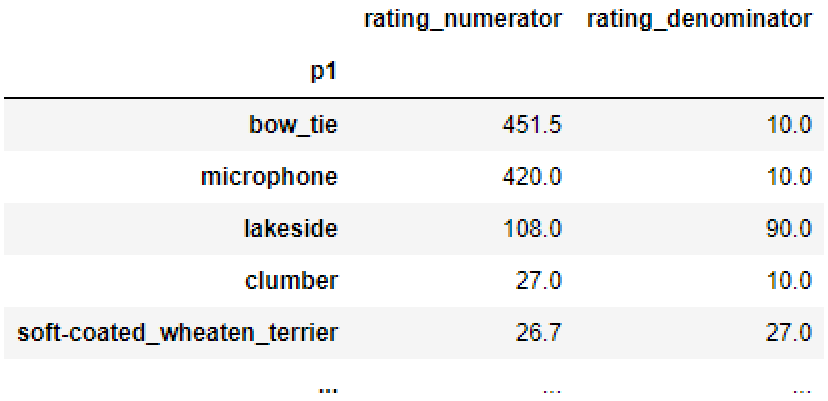


Figure 4 Top Ten Dog Breeds Bar Plot

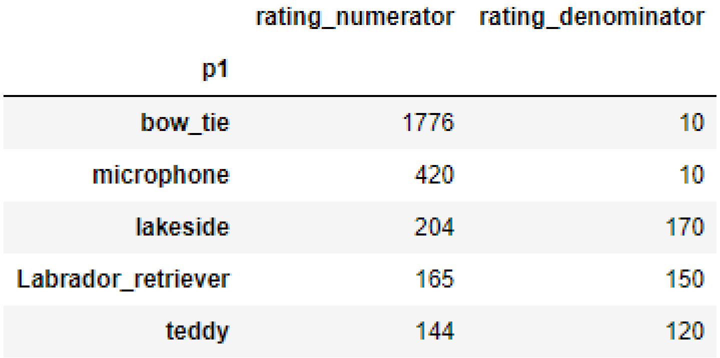
Q3: What are the most loved dog breeds?

When inspecting the highest average or maximum rated dog breeds, both statistics returned the highest values for images classified as **objects** (*bow tie*, *laptop*, etc.)



	rating_numerator	rating_denominator
p1		
bow_tie	451.5	10.0
microphone	420.0	10.0
lakeside	108.0	90.0
clumber	27.0	10.0
soft-coated_wheaten_terrier	26.7	27.0
...	...	...

Figure 5 Average Rating for the top predicted dog breed (p1)



p1	rating_numerator	rating_denominator
bow_tie	1776	10
microphone	420	10
lakeside	204	170
Labrador_retriever	165	150
teddy	144	120

Figure 6 Maximum Ratings for the top predicted dog breed (p1)

Which shows that the highest rated actual dog breed is still a *Labrador retriever.*

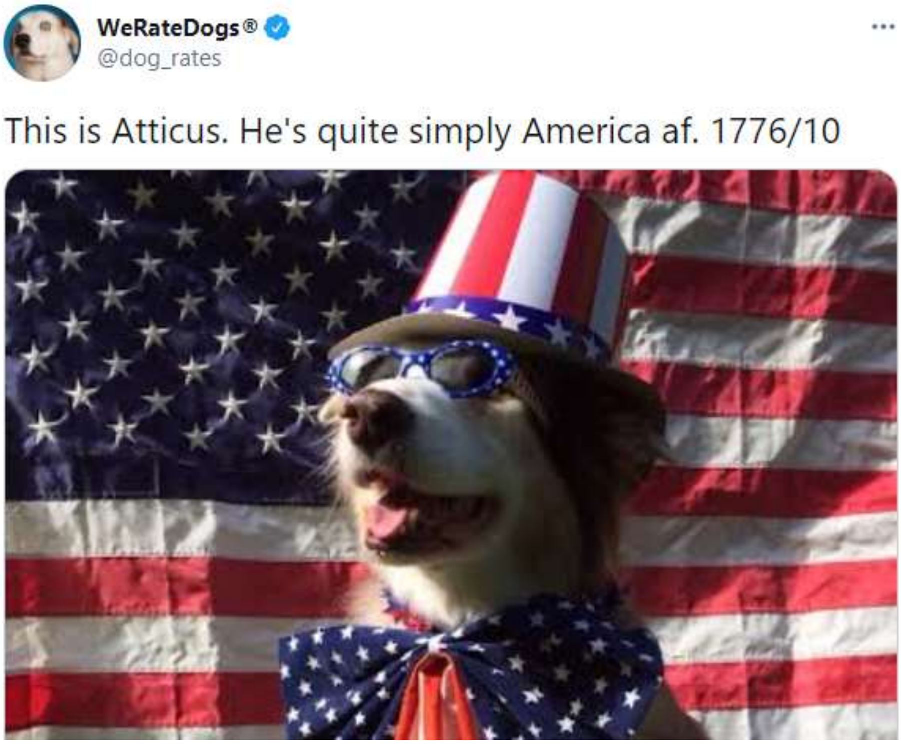
However, we need to investigate more the accuracy of the image predictions, and did the *WeRateDogs* account really give the highest ratings to objects or creatures other than dogs?!!

Hence, the following question:

Q4: How accurate are the dog breed classifications?

By inspecting the image that had the highest rating `(1776/10)`, which was classified as a `'bow_tie'`!

It turns out that the image is actually for a dog, a very cute one indeed!



WeRateDogs®
@dog_rates

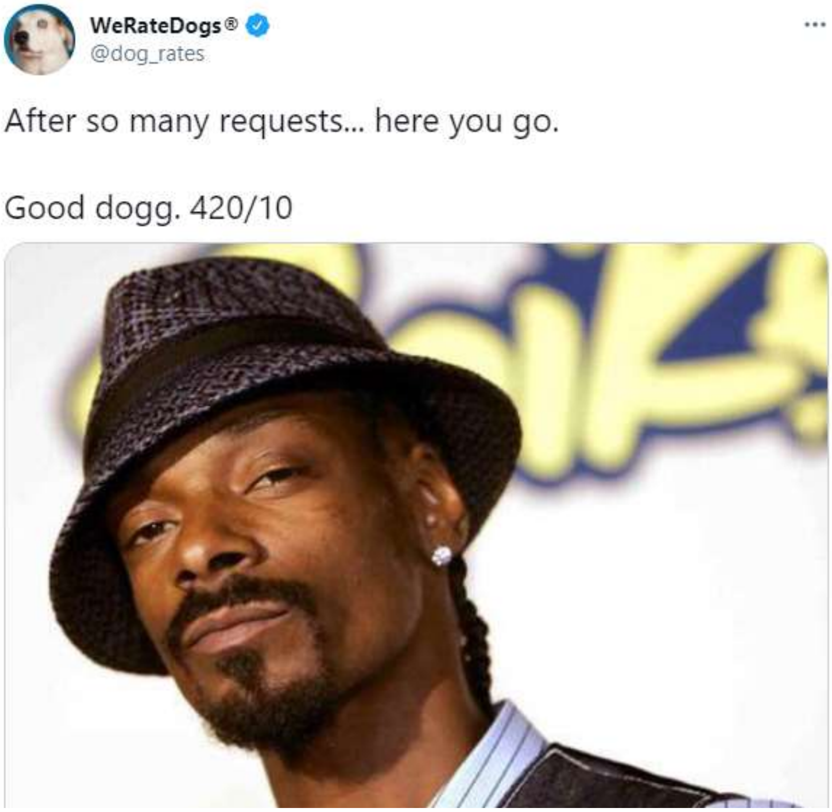
This is Atticus. He's quite simply America af. 1776/10

Figure 7 The Highest Rated dog Tweet

Another fun fact about this rating, is that the year `1776` is celebrated in the United States as the official beginning of the nation with the Declaration of Independence issued on July 4. *(Source: Wikipedia)* Which makes the rating perfectly suitable for the above picture.

Then, by inspecting the second highest rated tweet `(420/10)`, which was classified as `‘microphone’`.

This time the image is for ‘Snoop Dogg’ as shown in Figure 8, and not an actual dog, but the image prediction algorithm is still off, as he can’t be a microphone!



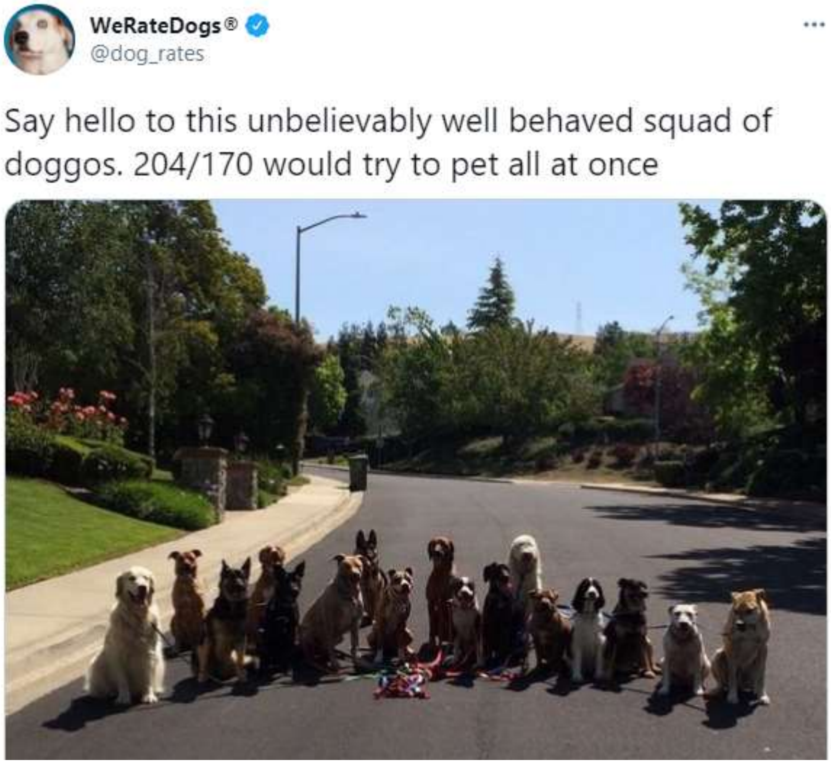
WeRateDogs®
@dog_rates

After so many requests... here you go.

Good dogg. 420/10

Figure 8 The Second Highest Rated Tweet

Getting to the third highest rated tweet `(204/170)`, it turns out to be for a group of dogs. But again the algorithm incorrectly classified them as `‘lakeside’`.



WeRateDogs®
@dog_rates

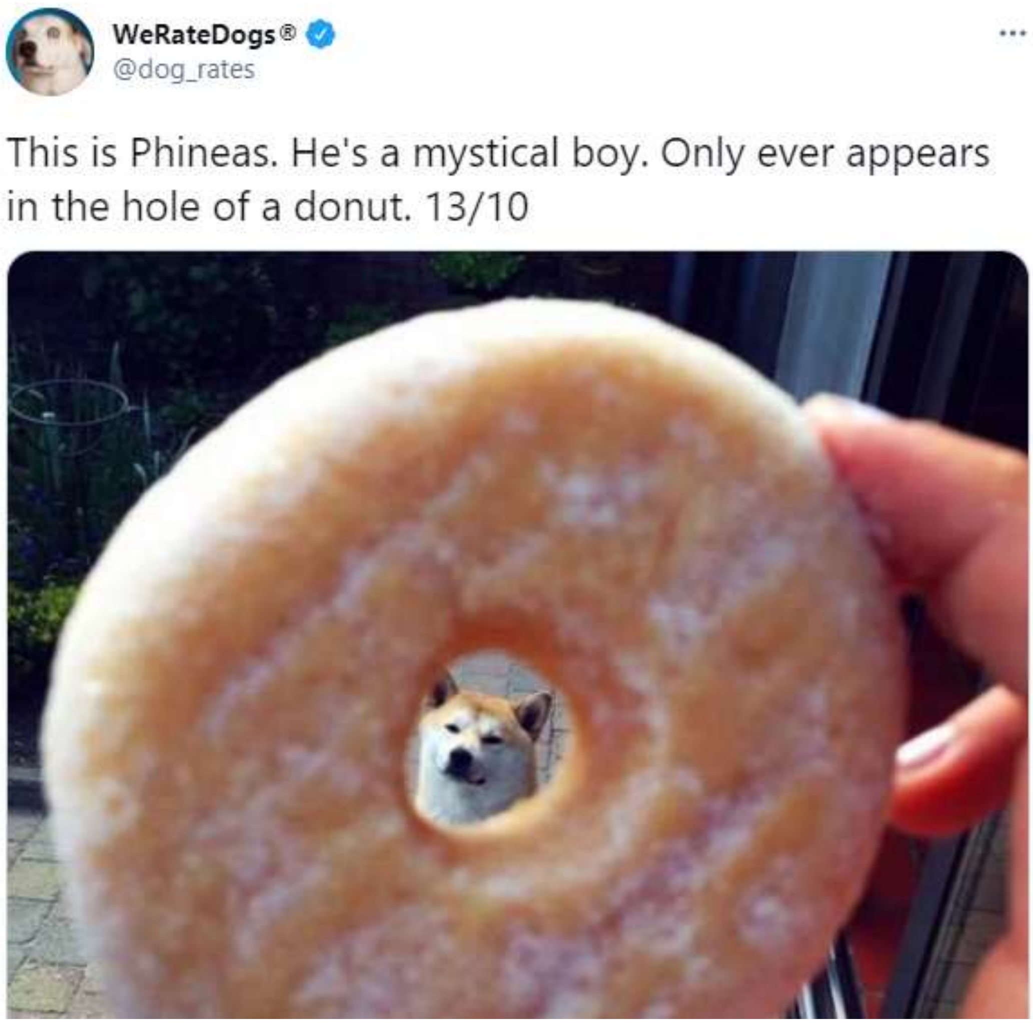
Say hello to this unbelievably well behaved squad of doggos. 204/170 would try to pet all at once

Figure 9 The Third Highest Rated Tweet

The previous rating also shows that the rating denominator was chosen by multiplying the number of dogs in the picture by 10 (=170), which will make the ratings for group dog pictures much higher than single ones.

So, it seems that the image prediction algorithm gets confused a lot, especially if there are other dominant objects in the image along with dogs.

Like the following image (Figure 10), which was incorrectly classified as an *orange*, due to the presence of a donut in front of the dog!



WeRateDogs®
@dog_rates

This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10

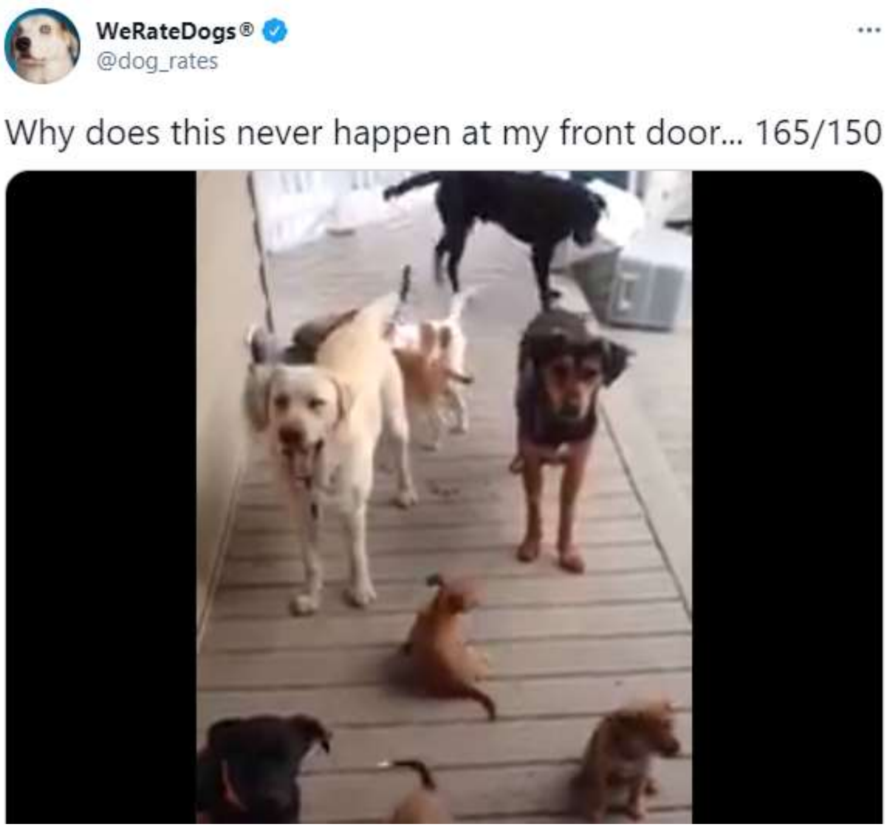
Figure 10 Incorrect Image Classification Example

Hence, we cannot totally rely on the image prediction dog breed variable `(p1)` for our analysis.

Q5: What does the abnormal high ratings represent?

Let's further investigate the very high ratings given to the highest correctly classified dog breed `‘Labrador retriever’`.

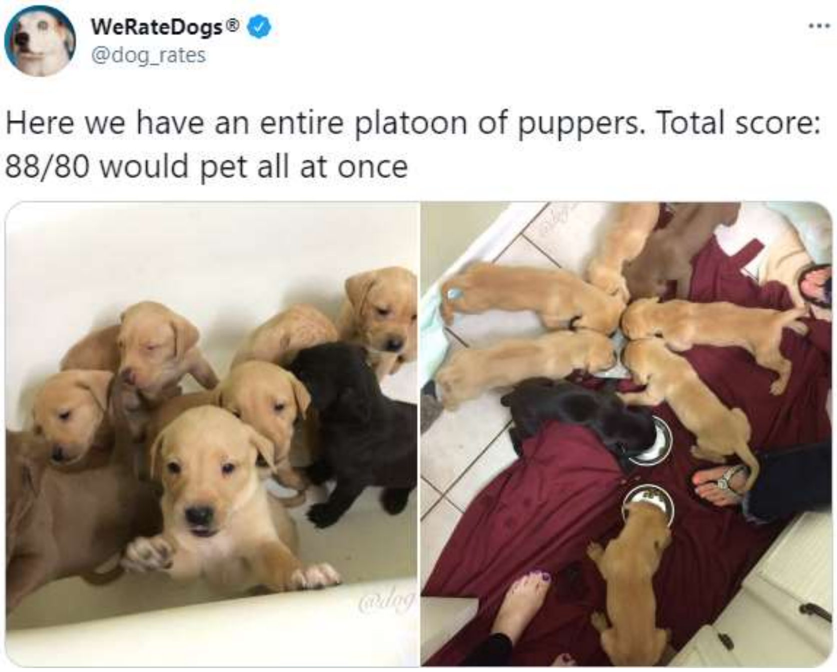
By inspecting the image that had the highest rating of `(165/150)`, which is correctly classified as a `'Labrador_retriever'`.



WeRateDogs®
@dog_rates

Why does this never happen at my front door... 165/150

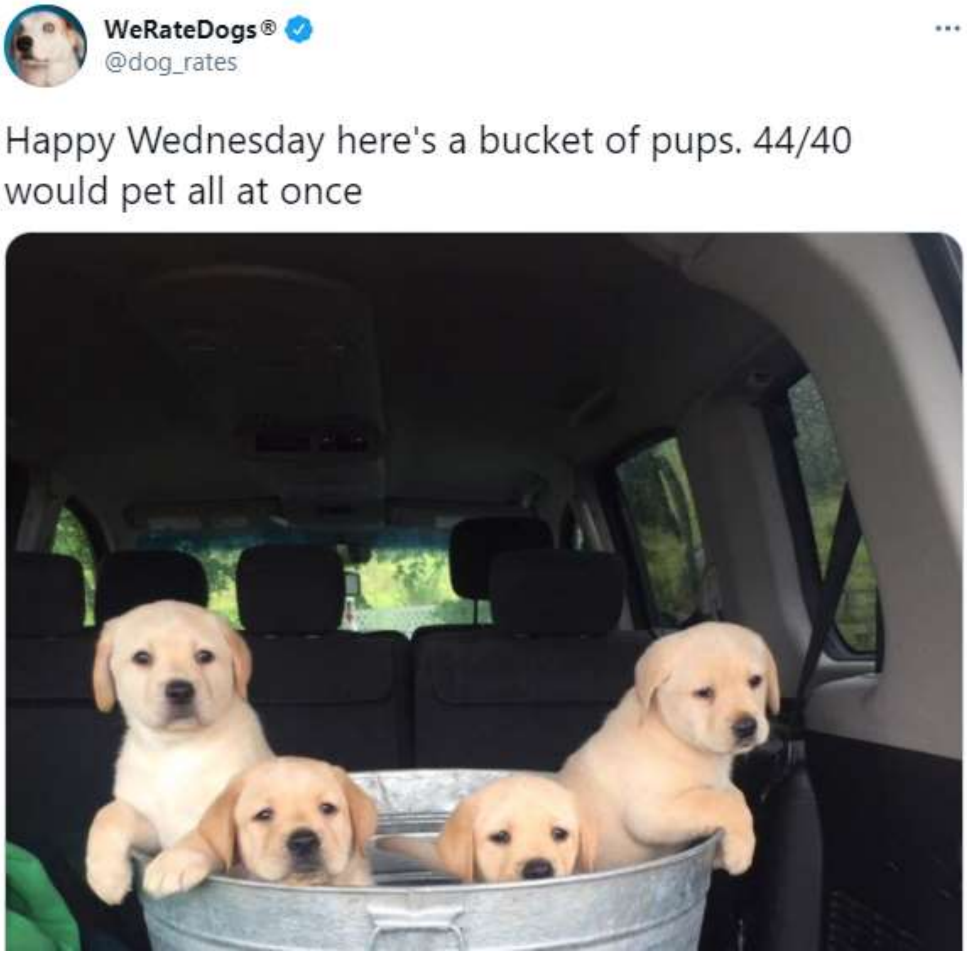
Followed by this one, which had a rating of `(88/80)`.



WeRateDogs®
@dog_rates

Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once

And the third place also goes to a group of cute puppies, with a rating of `(44/40)`.



WeRateDogs®
@dog_rates

Happy Wednesday here's a bucket of pups. 44/40 would pet all at once

Therefore, it seems that the highest ratings mostly goes to images that has a group of puppies.

Let’s check that assumption, by investigating the dog stages.

Q6: What are the most popular Dog Stages?

Indeed the most frequent tweets are about *puppers* (small doggos, AKA a Puppy), as shown in the following figure.

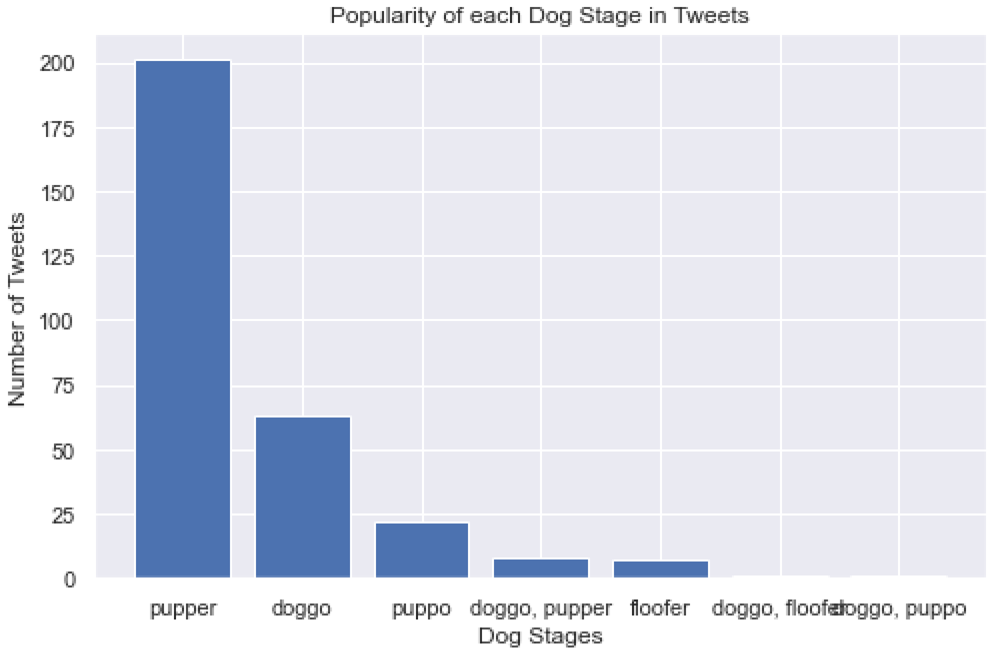


Figure 11 Puppers are the most popular in Tweets

And the highest retweet and favorite counts goes to images that has a group of doggos and puppos together, as shown in the following figure:

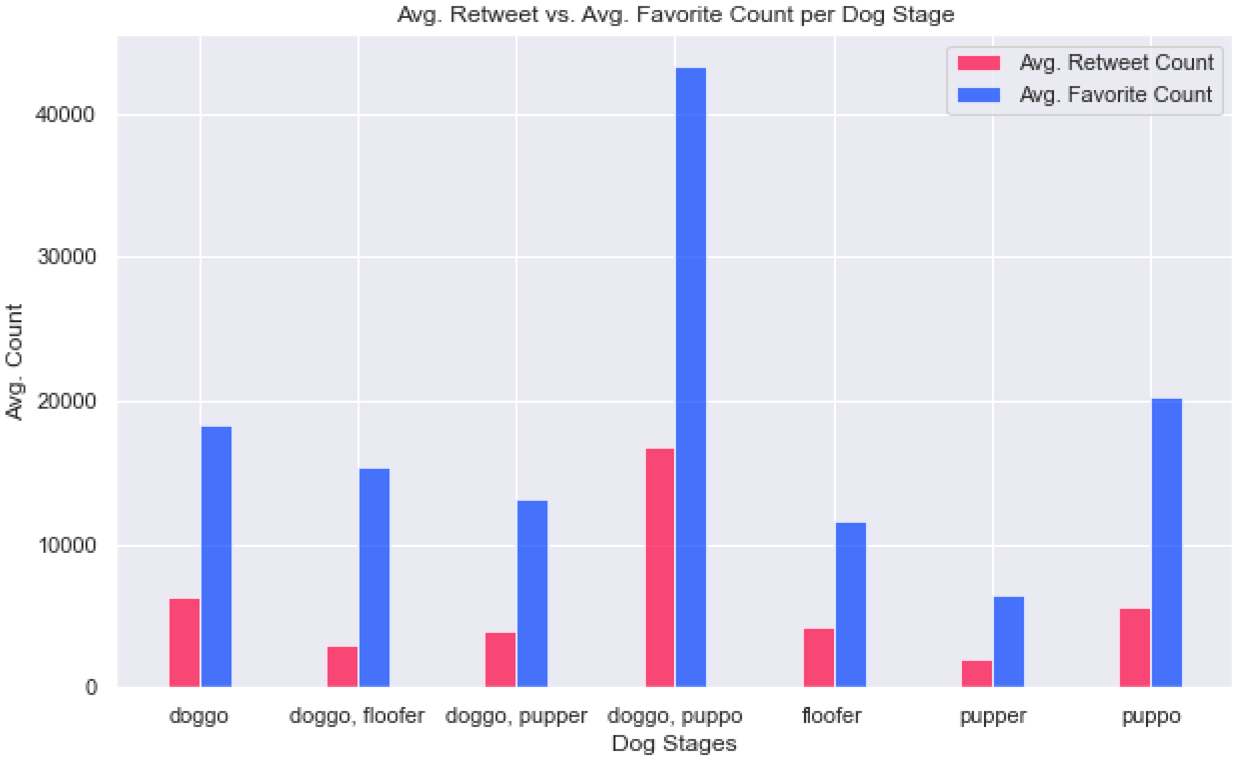


Figure 12 doggos and puppos together are the most retweeted & favorited