Data Insight and Visualization Report

~ Insights and Visualization from Wrangled Twitter Data ~

Audrey S Tan

April, 2019

Overview

With the cleansed dataset created from gathering twitter data pertaining to the popular WeRateDogs dog [ref 1 (https://en.wikipedia.org/wiki/WeRateDogs)] rating provider on Twitter, I analyzed and produced the following insights and visualizations:

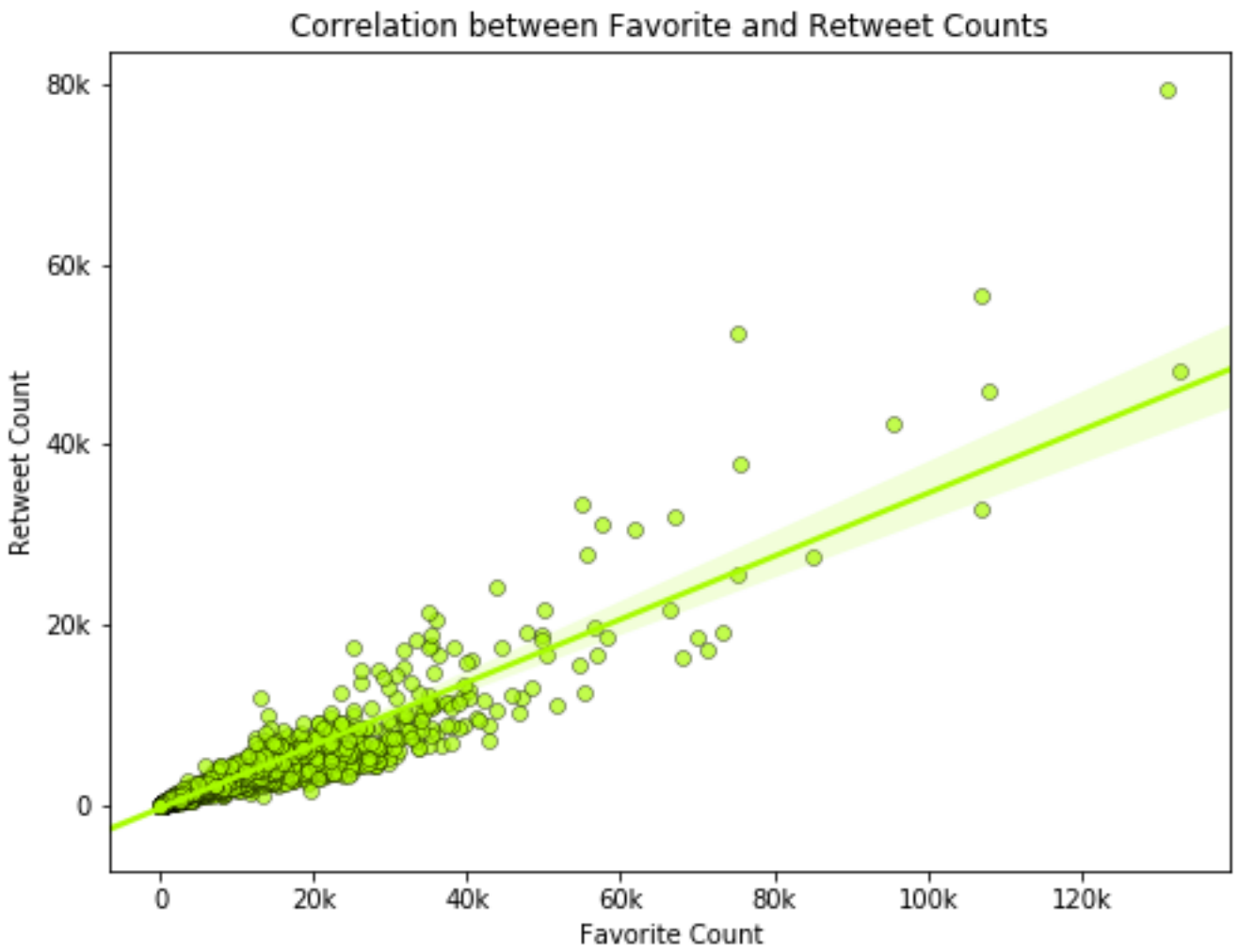
1) correlation between favorite and retweet counts.

2) the trend of favorite and retweet counts with respect to time and classification of dog species.

3) performance of the dog image classifier.

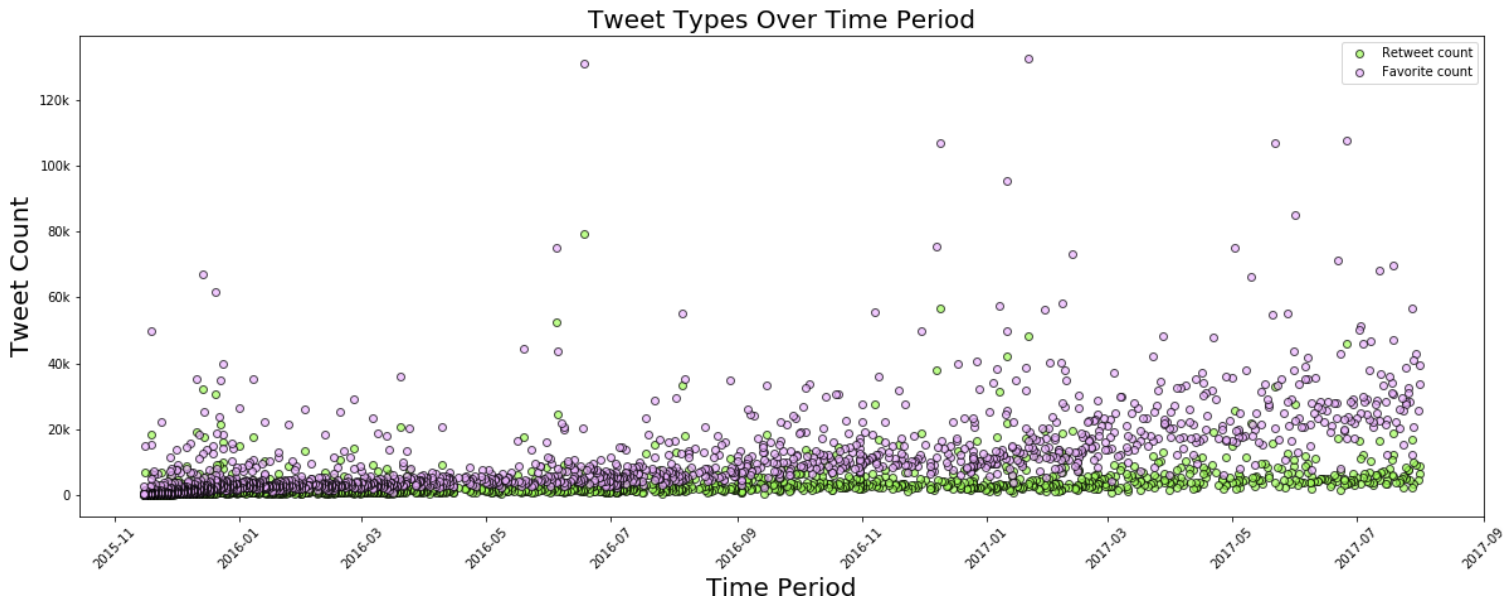
Correlation between favorite counts and retweet counts

favorite counts and retweet counts have a positive correlation. The intensity of the counts is heavily concentrated from the begining up to 40k favorite counts, with the counts largely below the regression line.



Favorite and retweet counts trends with respect to time period (Nov 2015 to Aug 2017)

From the begining till 2016-04, the intensity of both favorite and retweet counts is similar, although favorite counts are higher than retweet counts. This trend is even more conspicuous with the progression of time, with favorite counts steadily rising above retweet counts from around 2016-09. Interestingly, the bulk of retweet counts remains below 10k.

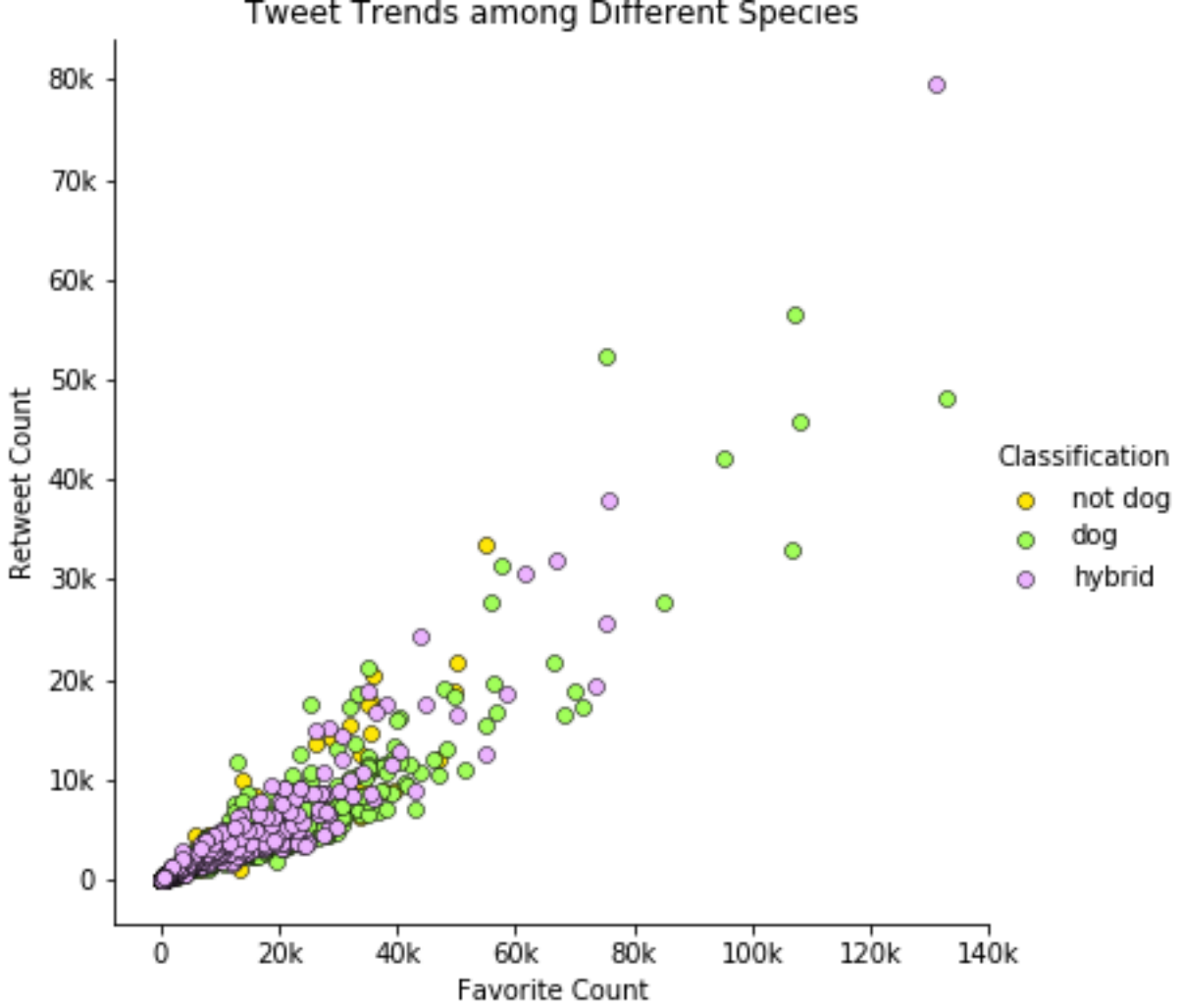


Favorite and retweet counts trends with respect to dog species by the image prediction classifier

Among the three speicies, both favorite and retweet counts have a positive correlation. The bulk of tweets is mainly from the species of `dog` and `hybrid` , which is in line with their respective species counts - 1194 dog and 472 hybrid.

```
# get a count of each species
df.p_class.value_counts()
```

```
dog        1194
hybrid      472
not dog     305
Name: p_class, dtype: int64
```



Given WeRateDogs is all about dogs, it is unconvincing there are tweets about `hybrid` and `not dog` species classified by the neural network dog breed classifier, which led us to take a look at the performance of the classifier next.

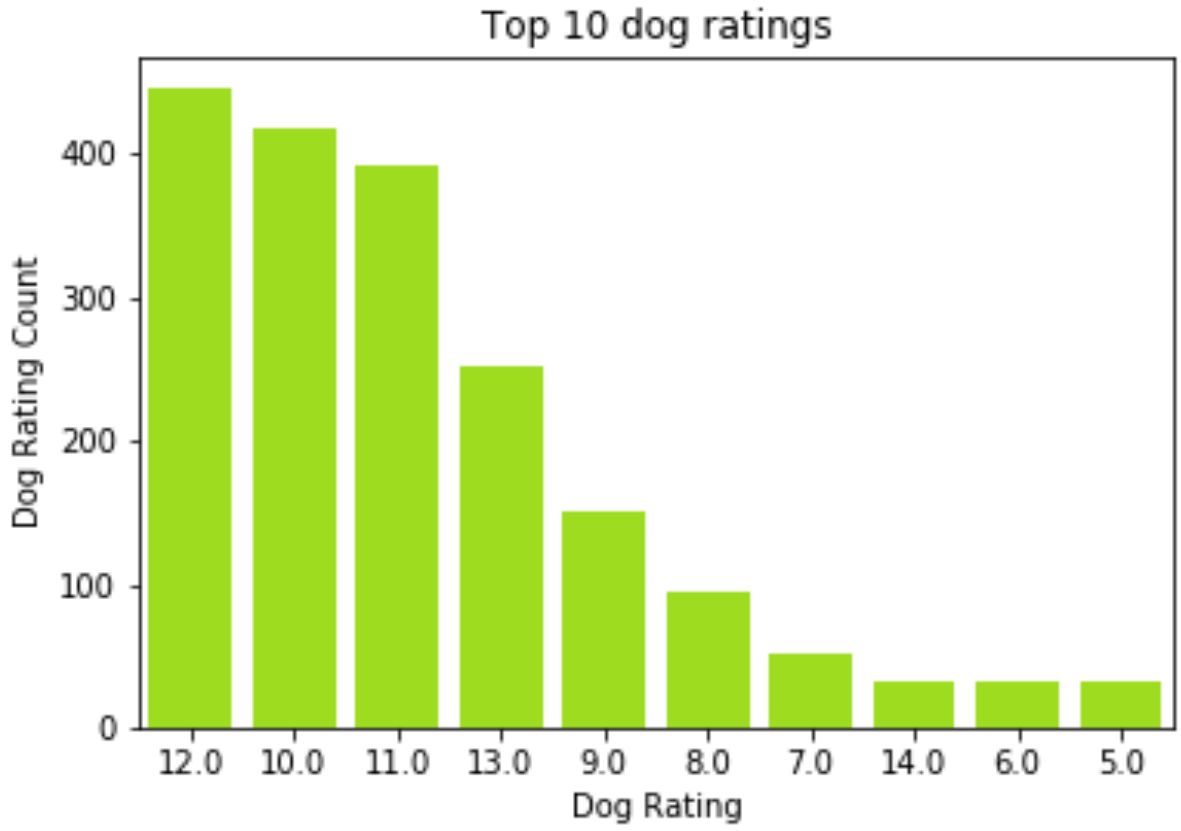
Performance of the dog image classifer

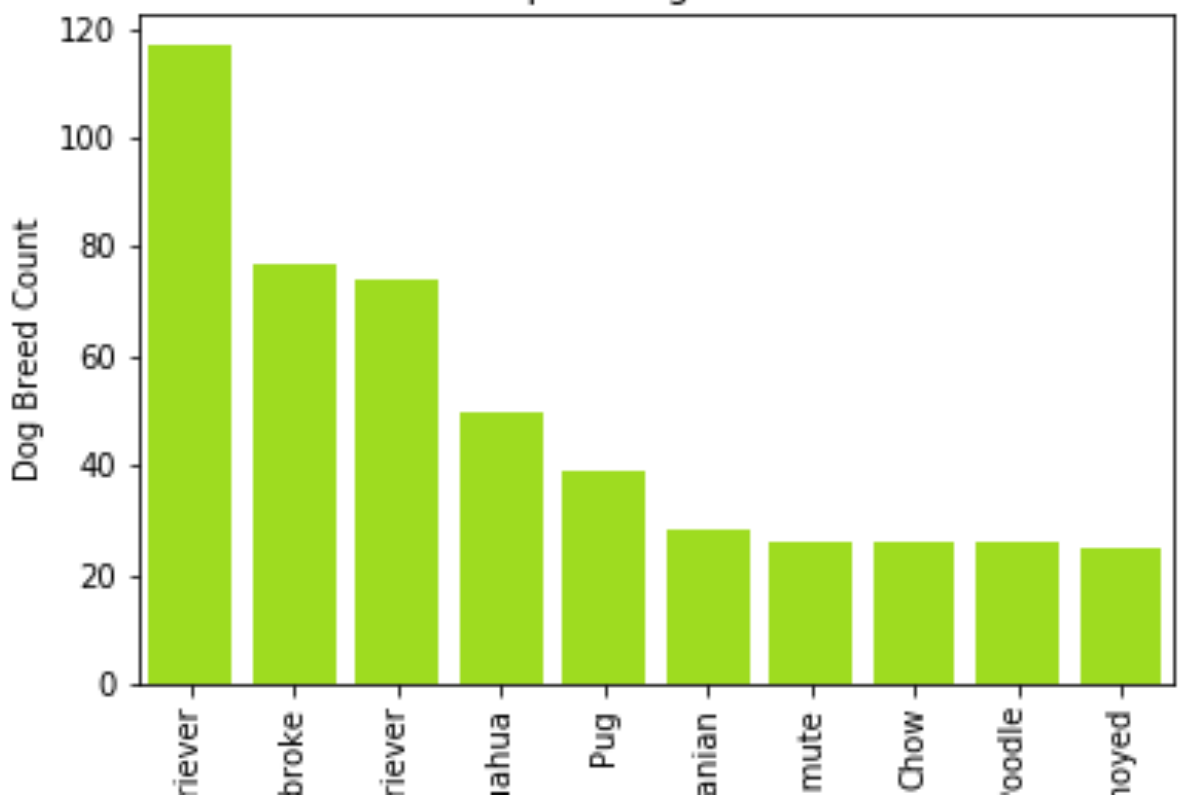
First, gathered some descriptive statistics on rating, favorite, retweet counts and the top 3 model confident predictions.

```
df[['rating_numerator','rating_denominator','retwt_cnt','fav_cnt','p1_conf','p2_conf','p3_conf','p_class']].describe()
```

	rating_numerator	rating_denominator	retwt_cnt	fav_cnt	p1_conf	p2_conf	p3_conf
count	1971.000000	1971.000000	1971.000000	1971.000000	1971.000000	1.971000e+03	1.971000e+03
mean	12.178985	10.477423	2784.449518	8949.106545	0.594558	1.345850e-01	6.016556e-02
std	41.607230	6.853275	4697.662893	12267.799790	0.272126	1.010527e-01	5.094156e-02
min	0.000000	2.000000	16.000000	81.000000	0.044333	1.011300e-08	1.740170e-10
25%	10.000000	10.000000	628.500000	1997.000000	0.363091	5.339800e-02	1.608055e-02
50%	11.000000	10.000000	1367.000000	4147.000000	0.587764	1.173970e-01	4.944380e-02
75%	12.000000	10.000000	3239.000000	11402.500000	0.847827	1.955655e-01	9.153815e-02
max	1776.000000	170.000000	79515.000000	132810.000000	1.000000	4.880140e-01	2.734190e-01

Next, take a look at the top 10 dog ratings and breeds





From the descriptive statistics, we checked the recipients of the highest

> ***favorite counts 132810.0***
>
> ***retweet counts 79515.0***
>
> ***p1 prediction confidence of 1.0***

then got their pictures from Twitter site.

The highest favorite counts recipient is:

```
#highest favorite counts
df[df.fav_cnt == 132810.0] # it's a dog (Labrador Retriever) with no name and a p1 prediction 0.196015
```

text	rating_numerator	rating_denominator	name	dog_stage	retwt_cnt	fav_cnt		jpg_url	...	p1	p1_conf
Here's a super supportive puppo participating ...	13.0	10.0	None	puppo	48265	132810	https://pbs.twimg.com/media/C2tugXLXgAArJO4.jpg		...	Lakeland Terrier	0.196015

It's a dog (Labrador Retriever) with no name and a p1 prediction 0.196015 and looks like this ...

The highest retweet counts recipient is:

```
#highest retweet counts
df[df.retwt_cnt == 79515.0] # It's a hybrid (Labrador Retriever) with no name and a p1 prediction 0.825333
```

text	rating_numerator	rating_denominator	name	dog_stage	retwt_cnt	fav_cnt	jpg_url	...	p1	p1_conf	p1_
Here's a doggo realizing you can stand in a po...	13.0	10.0	None	doggo	79515	131075	https://pbs.twimg.com/ext_tw_video_thumb/74423...	...	Labrador Retriever	0.825333	

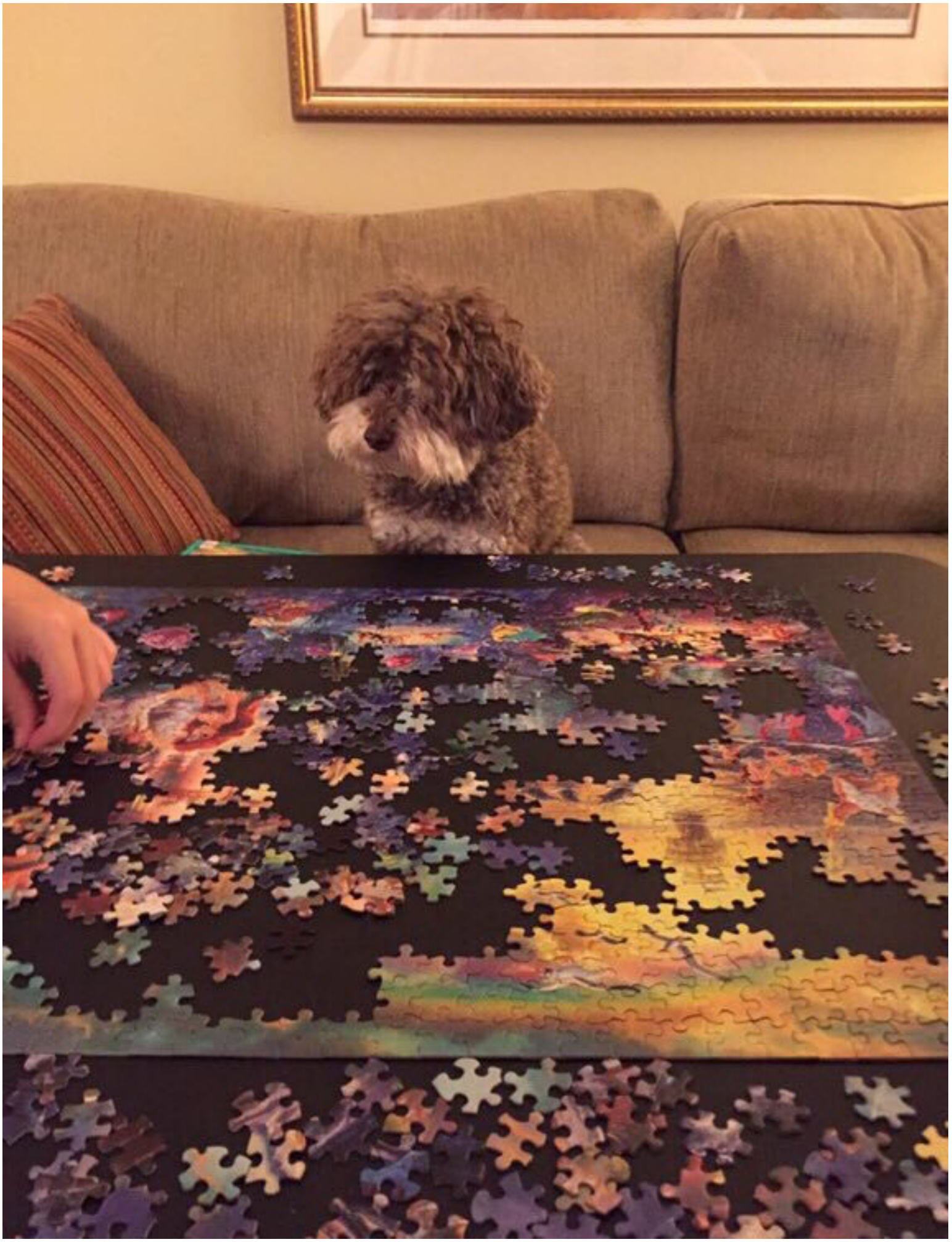
It's a hybrid (Labrador Retriever) with no name and a p1 prediction 0.825333 and looks like this ...

The highest p1 prediction confidence (1.0) recipient is:

```
# highest p1_conf
df[df.p1_conf == 1.0] #It's `not dog` named Shaggy (Spanish Water Dog) and a p1 prediction 1.0
```

	text	rating_numerator	rating_denominator	name	dog_stage	retwt_cnt	fav_cnt	jpg_url	...	p1	p1_conf	p
1	This is Shaggy. He knows exactly how to solve ...	10.0	10.0	Shaggy	None	1110	3172	https://pbs.twimg.com/media/CUS9PlUWwAANeAD.jpg	...	Jigsaw Puzzle	1.0	

It's `not dog` named Shaggy (Spanish Water Dog) and a p1 prediction 1.0 and looks like this ...



So what is performance of the image classifier afterall ?

Well, the `dog` and `hybrid` share the same rating of 13 while the `not dog` has a rating of 10, which rank 4th and 2nd respectively in the top 10 dog ratings. Both the `dog` and `hybrid` are actually Labrador Retriever which ranks 3rd in the top 10 dog breeds. The `not dog` was misclassified but is actually a Spanish Water Dog, though the breed is not among the top 10 dog breeds.

So of the 1971 entries in the `twitter_archive_master` dataset, 305 entries were misclassified as `not dog`, 472 as `hybrid` (i.e. `may be dog`). Only 1194 were correctly classified as `dog` yet none attains the highest p1 predection of 1. Ironically, the `not dog` has the highest prediction confidence of 1.

The numbers represent a 60.58% (=1194/1971 x 100%) chance the model correctly identifies a dog as a dog. Obviously, the model has more room for improvement.