act_report

February 17, 2019

0.1 Act_Report

Starting with around five thousand rows of data and after the cleaning process ending with around three hundred is a significant drop off. This is sometimes the case though when there is a lot of missing data or very unclean data (where we lost approximately three thousand rows) and data needs to be merged together that is incomplete from one set to another. With the end result being a completely filled dataframe in mind, dropping so many observations was the only way to accomplish this.

[1] Looking at any numerical features in relation to one another, it's obvious there isn't a whole lot of correlation going on between much of the data. Several features offered nothing in regard to insight, like 'breed_confidence', where almost every relationship it has was a non-descript distrobution. The same can be said for the most part for 'tweet_image_number' and 'normalized_rating'.

[2] No particular dog 'stage' was incredably more popular than another in regards to generating retweets. We can see in the graph below, which compares observations of a particular 'stage' vs the rest of th retweets (labeled None) for the four catagories. 'doggo' has slightly more than the others, but nothing crazy. Not to be confused, because a handful of observations had multiple 'stage' tags I didn't want to stack these together, so each bar is representative of the entire dataframe, targetting only one 'stage' tag.

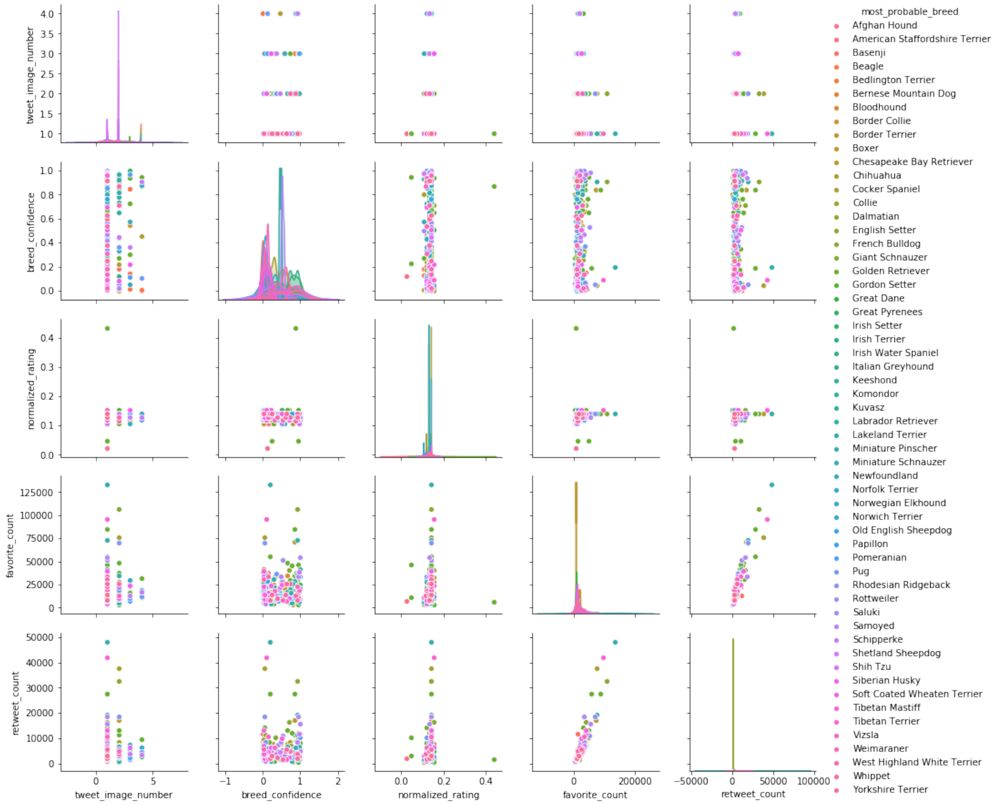
[3] The two features with the greatest correlation, which is no surprise, is 'favorite_count' and 'retweet_count'. It just makes sense that these two go hand in hand as we can tell from this scatter plot with a regression line. Personally I like this graph from Seaborn because besides showing the data points, it plots a regression line and shows the distrobution of both features against each axis.

[4] Another (slightly) positive relationship which makes sense is the relationship between 'breed_confidence' and 'tweet_image_number'. I would image that the more certain the algorithm is in making a prediction, the more likely it would be that it would come to this conclusion on the first image. This could be due to several factors; maybe the first image was a higher quality photo, maybe it was a clearer photo, maybe that particular breed is more easily identified. No matter the reason tho it does make sense to see this.

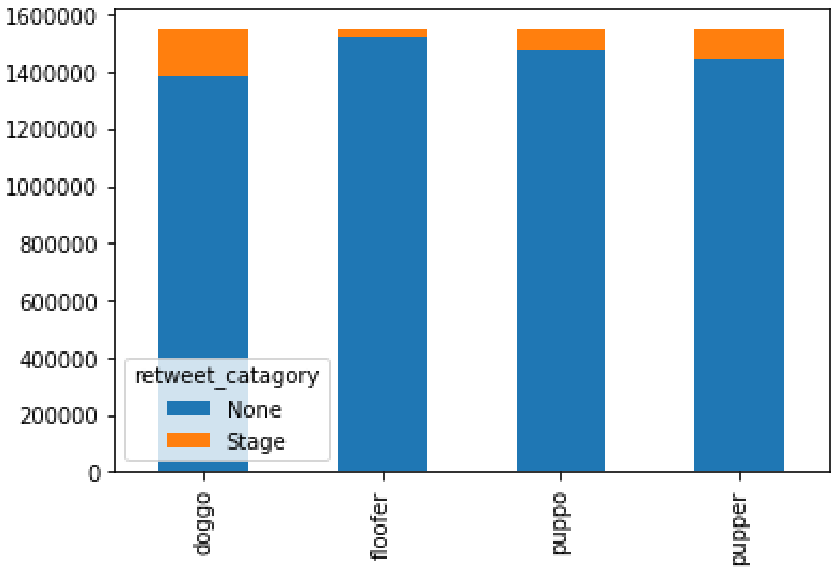
[5 & 6] The next two graphs show some more positive correlations, which again if we think about it is not surprising. The higher the 'normalized_rating' gets, the more retweets or favorites the dog got. This makes sense, and likely would have been even clearer should the data have been a little better (their rating system not so strange and the so many observations needing to be dropped). Note: some additional observations were dropped for these graphs to exaggerate the regression line. This would have been apparent in the data itself but some outliers prevented the correct distrobution.

[7] Another thing I looked at was the distrobution of breeds being selected during photos 1-4 of the tweets. While the majority of high confidence levels came during the furst photo, no particular breed was more prone to being selected later compared to the other breeds.

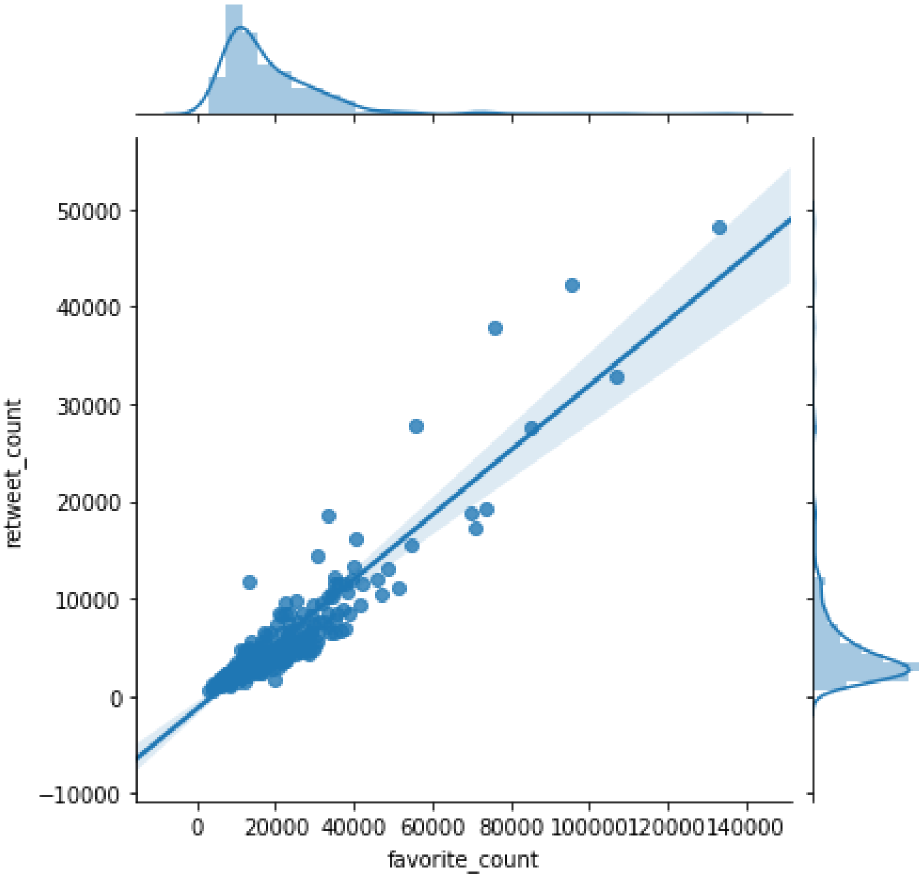
[8] With the limited time we had to explore these data sets and a relatively large amount of data being lost for various reasons I thought the insights that could be drawn from this activity were predictable. I wasn't particularly surprised by anything, although portions of the activity were good (worthwhile) like web scraping to aggregate the data. The last thing we'll look at is the popularity of an individual breed. Not surprising, America's favorite dog the Golden Retriever was by far the most retweeted in the group.



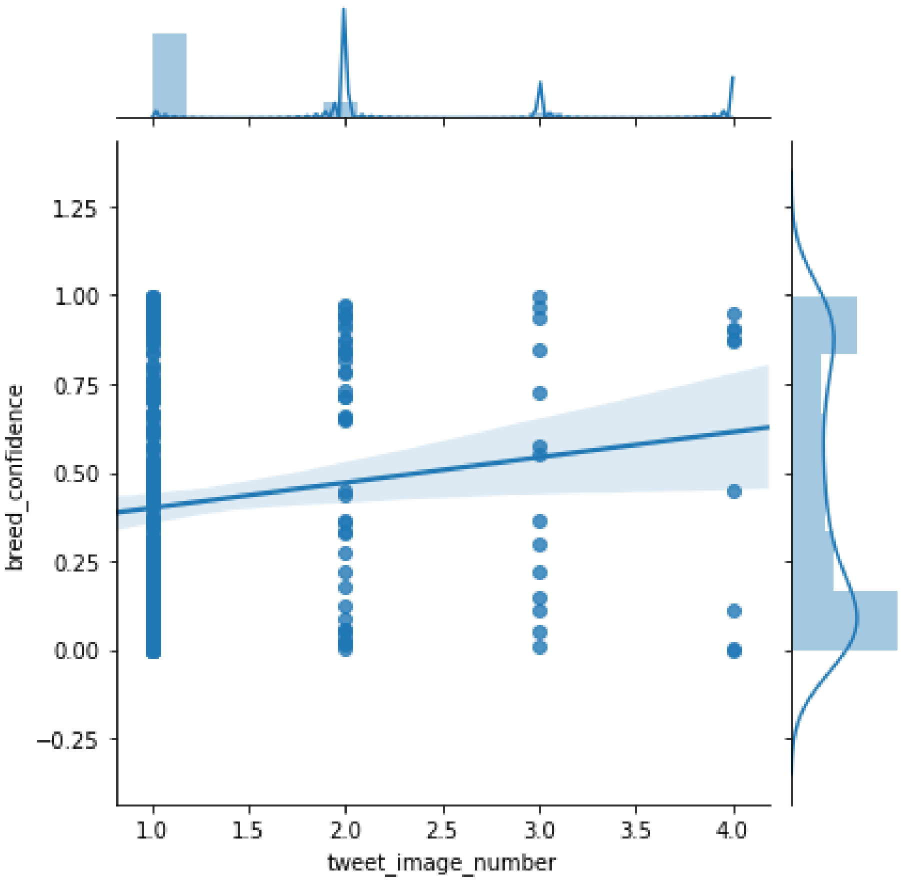
[1]


1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
retweet_catagory
None
Stage
doggo
floofer
puppo
pupper

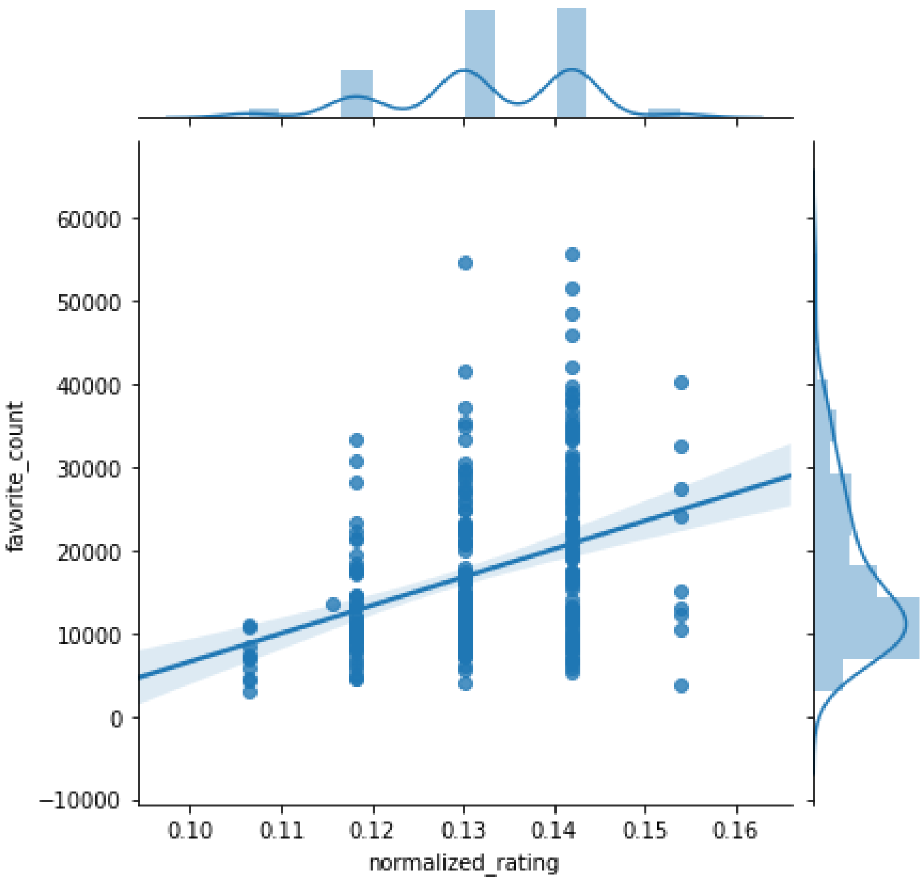
[2]


50000
40000
30000
20000
10000
0
−10000
retweet_count
0
20000
40000
60000
80000
100000
120000
140000
favorite_count

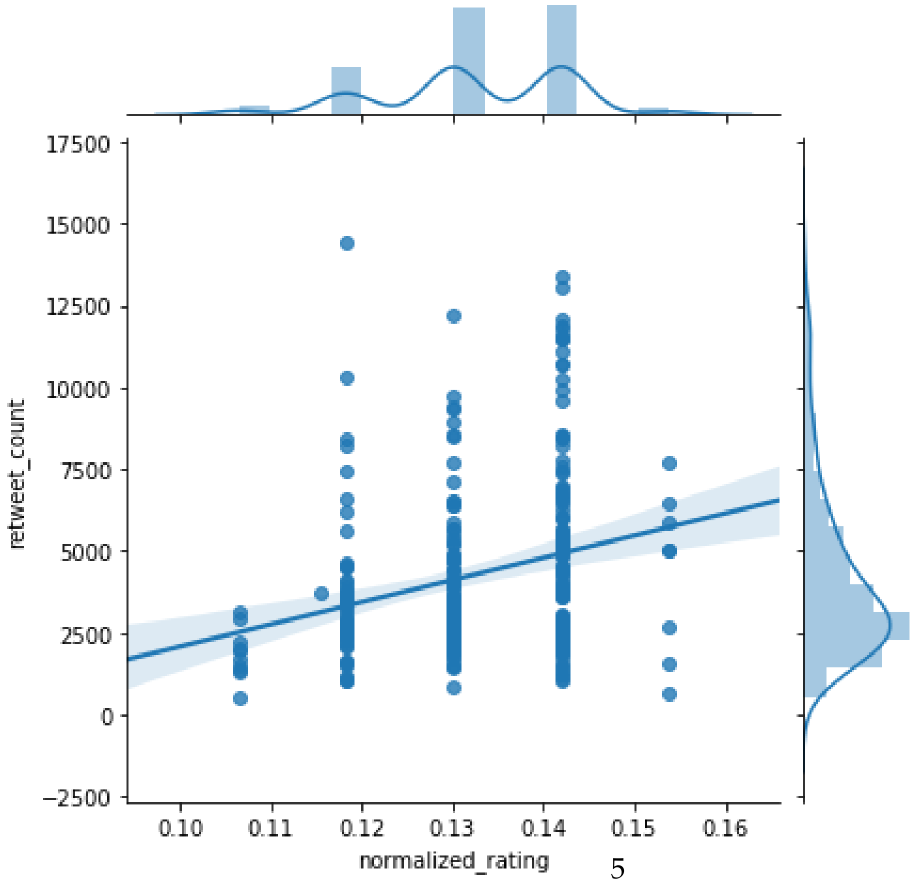
[3]

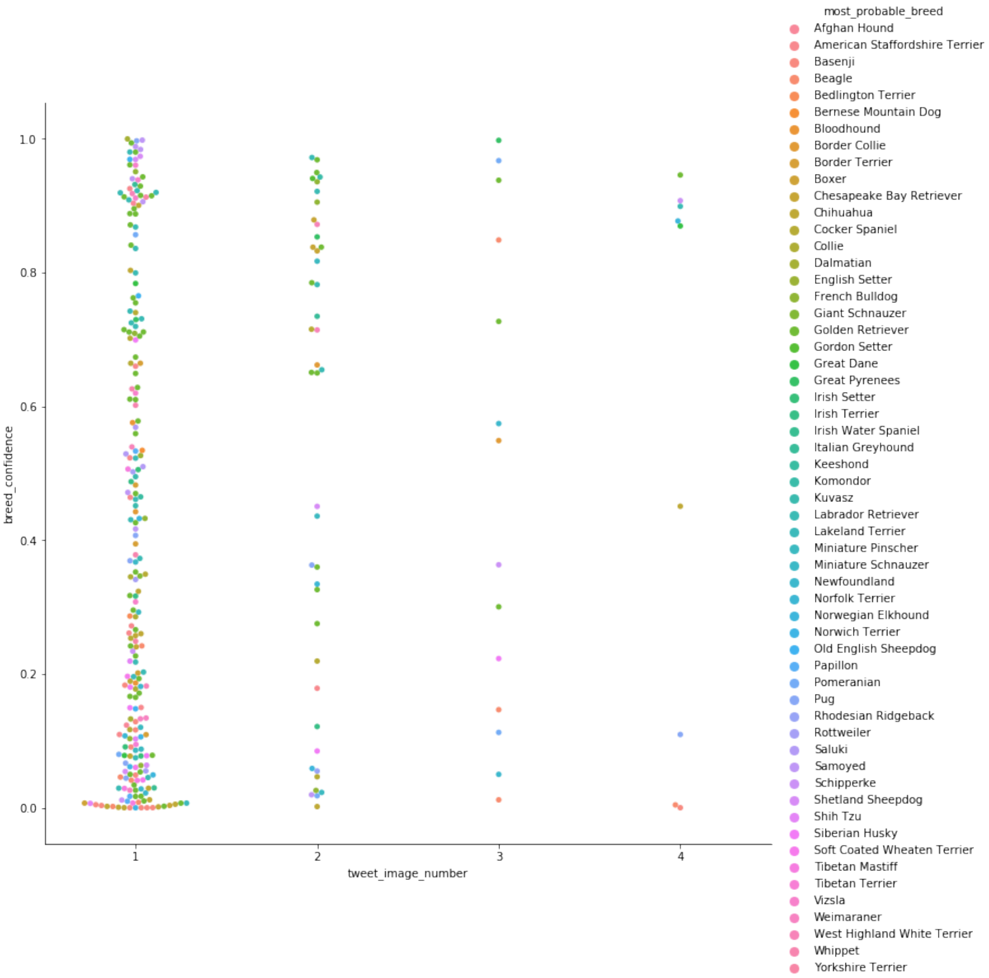

1.25
1.00
0.75
0.50
0.25
0.00
−0.25
breed_confidence
1.0
1.5
2.0
2.5
3.0
3.5
4.0
tweet_image_number

[4]

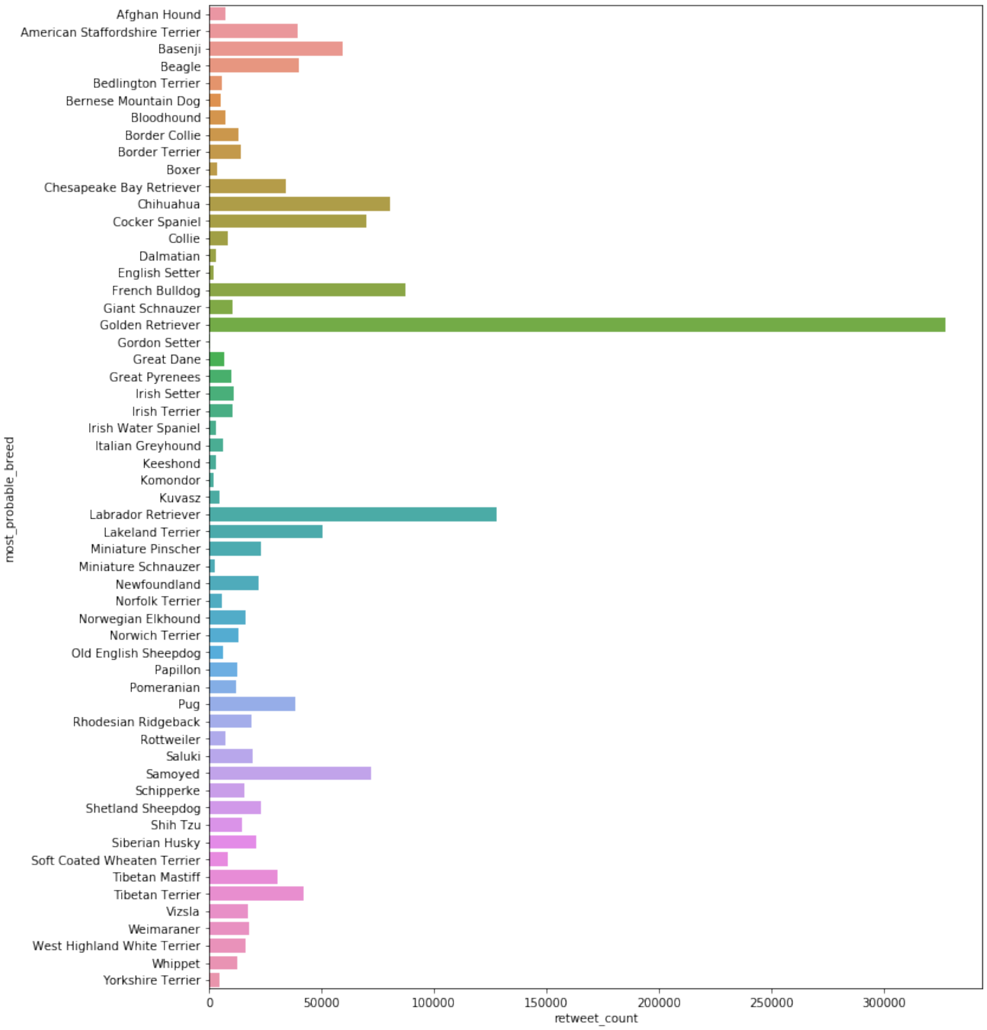

60000
50000
40000
30000
20000
10000
0
−10000
favorite_count
0.10
0.11
0.12
0.13
0.14
0.15
0.16
normalized_rating

[5 & 6]


17500
15000
12500
10000
7500
5000
2500
0
−2500
retweet_count
0.10
0.11
0.12
0.13
0.14
0.15
0.16
normalized_rating


most_probable_breed
Afghan Hound
American Staffordshire Terrier
Basenji
Beagle
Bedlington Terrier
Bernese Mountain Dog
Bloodhound
Border Collie
Border Terrier
Boxer
Chesapeake Bay Retriever
Chihuahua
Cocker Spaniel
Collie
Dalmatian
English Setter
French Bulldog
Giant Schnauzer
Golden Retriever
Gordon Setter
Great Dane
Great Pyrenees
Irish Setter
Irish Terrier
Irish Water Spaniel
Italian Greyhound
Keeshond
Komondor
Kuvasz
Labrador Retriever
Lakeland Terrier
Miniature Pinscher
Miniature Schnauzer
Newfoundland
Norfolk Terrier
Norwegian Elkhound
Norwich Terrier
Old English Sheepdog
Papillon
Pomeranian
Pug
Rhodesian Ridgeback
Rottweiler
Saluki
Samoyed
Schipperke
Shetland Sheepdog
Shih Tzu
Siberian Husky
Soft Coated Wheaten Terrier
Tibetan Mastiff
Tibetan Terrier
Vizsla
Weimaraner
West Highland White Terrier
Whippet
Yorkshire Terrier
breed_confidence
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
tweet_image_number

[7]


Afghan Hound
American Staffordshire Terrier
Basenji
Beagle
Bedlington Terrier
Bernese Mountain Dog
Bloodhound
Border Collie
Border Terrier
Boxer
Chesapeake Bay Retriever
Chihuahua
Cocker Spaniel
Collie
Dalmatian
English Setter
French Bulldog
Giant Schnauzer
Golden Retriever
Gordon Setter
Great Dane
Great Pyrenees
Irish Setter
Irish Terrier
Irish Water Spaniel
Italian Greyhound
Keeshond
Komondor
Kuvasz
Labrador Retriever
Lakeland Terrier
Miniature Pinscher
Miniature Schnauzer
Newfoundland
Norfolk Terrier
Norwegian Elkhound
Norwich Terrier
Old English Sheepdog
Papillon
Pomeranian
Pug
Rhodesian Ridgeback
Rottweiler
Saluki
Samoyed
Schipperke
Shetland Sheepdog
Shih Tzu
Siberian Husky
Soft Coated Wheaten Terrier
Tibetan Mastiff
Tibetan Terrier
Vizsla
Weimaraner
West Highland White Terrier
Whippet
Yorkshire Terrier
most_probable_breed
0
50000
100000
150000
200000
250000
300000
retweet_count

[8]