# Project: Wrangle and Analyze Data

# WeRateDogs Twitter Archive Analysis

## Introduction

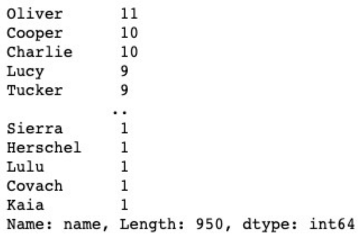
The WeRateDogs Twitter account is famous for giving rates and funny comments about dogs, summing over 8.8 million followers. It was created in 2016 and the rates refer to the review of pictures of dogs in different poses. These rates do not follow any specific metric and are often over 10/10.

The data analyzed in this project was provided by Udacity's servers and Twitter API and had information on 5000+ tweets. However, only 2356 tweets had ratings and those were the ones investigated. Moreover, during the data cleaning procedure, some tweets were removed, because they were considered to lack some meaningful information for the current analysis.

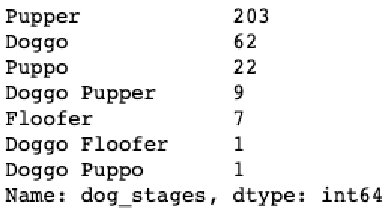
The final DataFrame used in the current analysis was composed of 1969 tweets and provided information about the tweet ID, the timestamp, source, text, URL, rating denominator, and numerator of the tweet. It also included information about the dog name and dog stage (if given), the dog breed  (predicted by a neural network), the URL for the dog image, and retweet and favorite counts.

## Insights

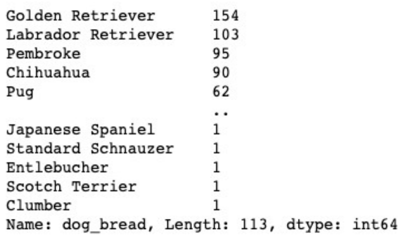
Analyzing the data, it is possible to observe that the most common name dog name is Oliver, the most common dog stage is pupper, appearing in 66.56% of dog stages citations, and the most common dog breed is Golden Retriever,  being detected in 9.27% of images by the breed prediction neural network.



```
Oliver      11
Cooper      10
Charlie     10
Lucy         9
Tucker       9
            ..
Sierra       1
Herschel     1
Lulu         1
Covach       1
Kaia         1
Name: name, Length: 950, dtype: int64
```

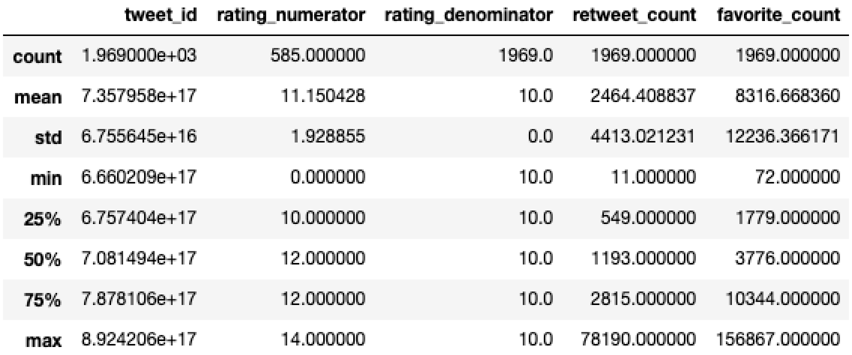


```
Pupper             203
Doggo              62
Puppo              22
Doggo Pupper       9
Floofer            7
Doggo Floofer      1
Doggo Puppo        1
Name: dog_stages, dtype: int64
```



```
Golden Retriever       154
Labrador Retriever     103
Pembroke               95
Chihuahua              90
Pug                    62
                       ..
Japanese Spaniel       1
Standard Schnauzer     1
Entlebucher            1
Scotch Terrier         1
Clumber                1
Name: dog_bread, Length: 113, dtype: int64
```

Moreover, it was possible to observe that the average retweet and favorite counts were respectively 2,464.41 and 8,316.66. They respectively had minimum values of 11 and 72, and maximum values of 78,190 and 156,867. Also, 75% of tweets have more than 1,779 favorites and 549 retweets. This information shows that there are tweets with really low interaction, but most of the tweets have a high number of retweets and favorites, and users prefer to favorite instead of a retweet.

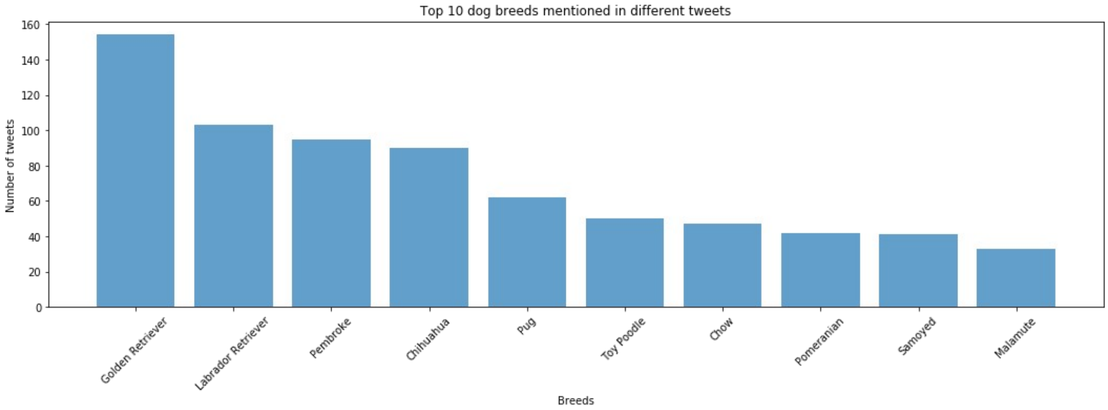


| | tweet_id | rating_numerator | rating_denominator | retweet_count | favorite_count |
|---|---|---|---|---|---|
| **count** | 1.969000e+03 | 585.000000 | 1969.0 | 1969.000000 | 1969.000000 |
| **mean** | 7.357958e+17 | 11.150428 | 10.0 | 2464.408837 | 8316.668360 |
| **std** | 6.755645e+16 | 1.928855 | 0.0 | 4413.021231 | 12236.366171 |
| **min** | 6.660209e+17 | 0.000000 | 10.0 | 11.000000 | 72.000000 |
| **25%** | 6.757404e+17 | 10.000000 | 10.0 | 549.000000 | 1779.000000 |
| **50%** | 7.081494e+17 | 12.000000 | 10.0 | 1193.000000 | 3776.000000 |
| **75%** | 7.878106e+17 | 12.000000 | 10.0 | 2815.000000 | 10344.000000 |
| **max** | 8.924206e+17 | 14.000000 | 10.0 | 78190.000000 | 156867.000000 |

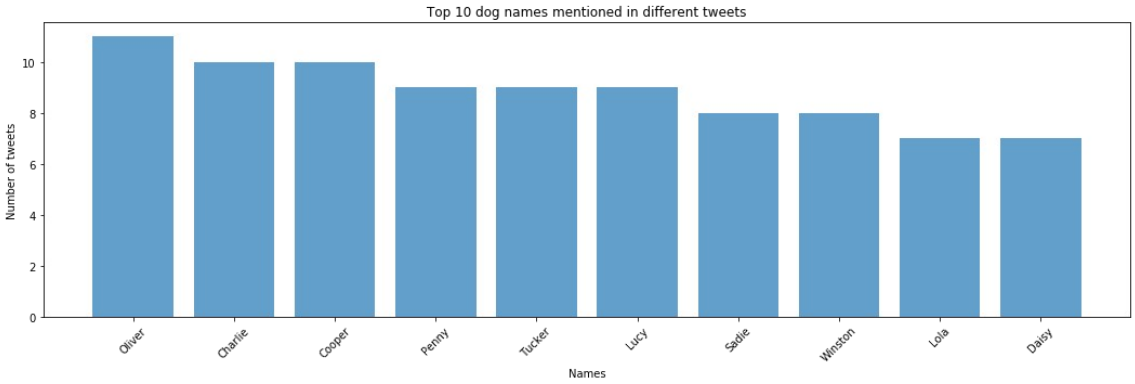
Tweets with a high favorite count usually have a dog identified as Golden Retriever, being this observed in 14.13% of the tweets with dog breed information. Also, tweets with a high favorite count usually have a dog stage classification of doggo, being this observed in 39.56% of the tweets with dog stage information.

## Visualization

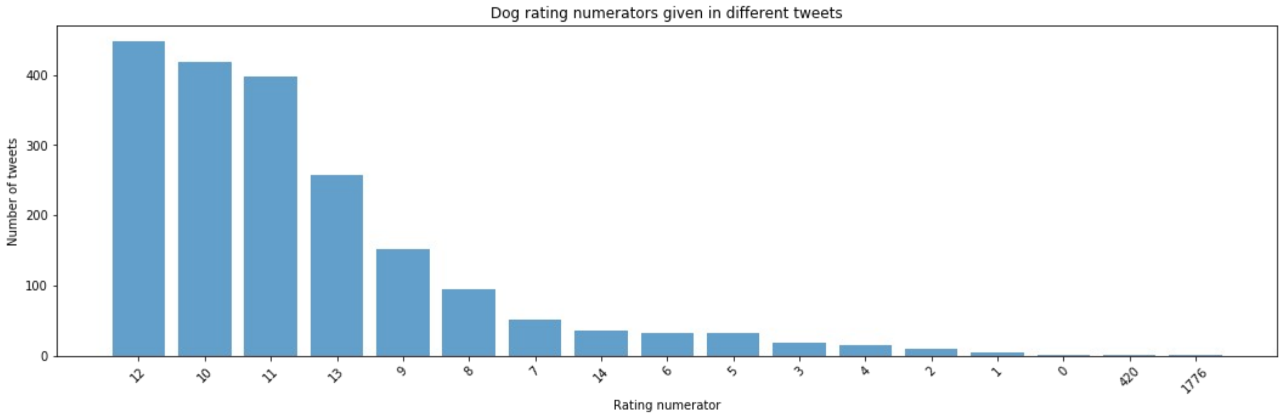
As mentioned in the insights, the most common dog breed is Golden Retriever, this is clearly presented in the following plot, which are also presented the top 10 most identified dog breed.



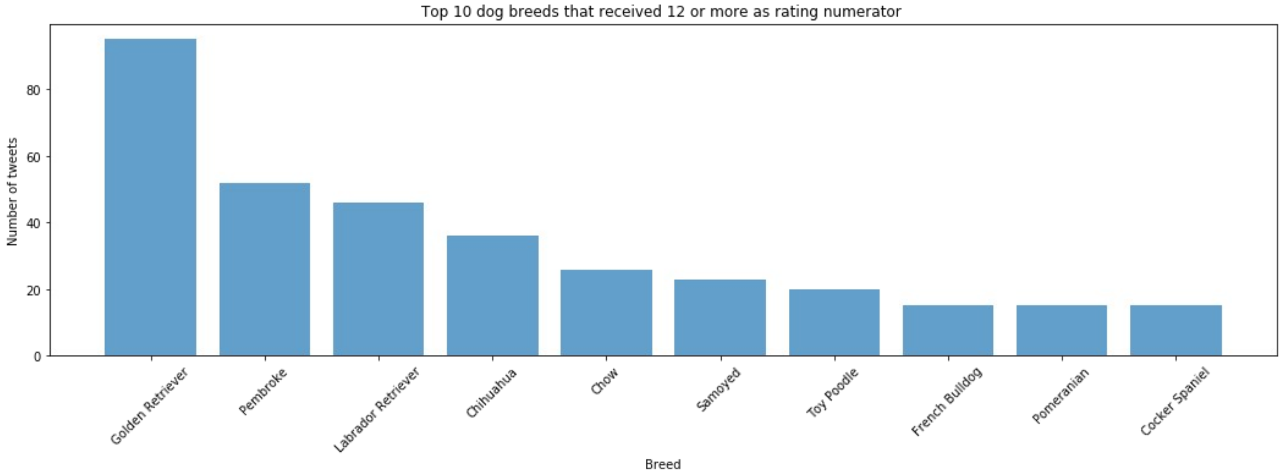
In the following plot, it is possible to observe the top 10 most mentioned dog names, being Oliver the most common name, accordingly with what was presented in the insights.



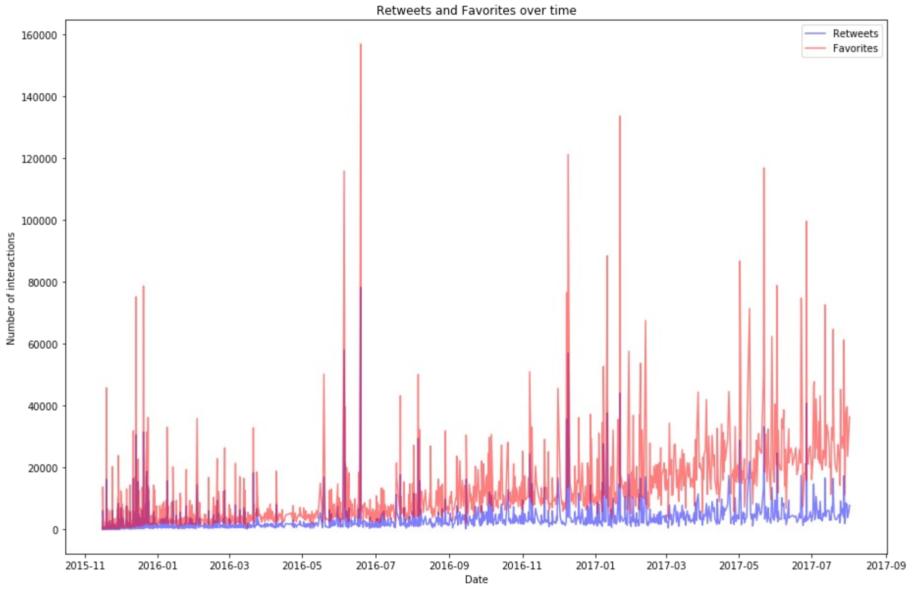
We can also draw some conclusions about the rating numerators. As can be observed in the following plot, the most common numerator is 12 and all the top 3 numerators are above 10, indicating that "they are good dogs".



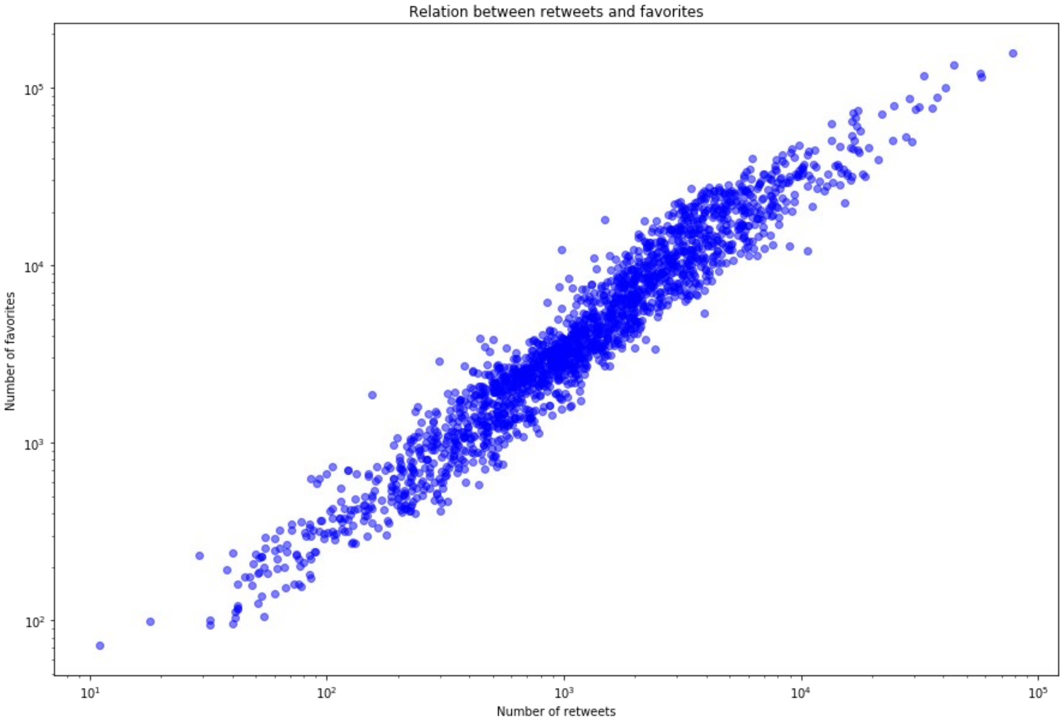
Another fact regarding dog numerators is the dog breed that is the most rated with high values. In the following plot, we can observe the top 10 dog breeds that were rated with 12 or more in the numerator, being the Golden Retriever the most common one.



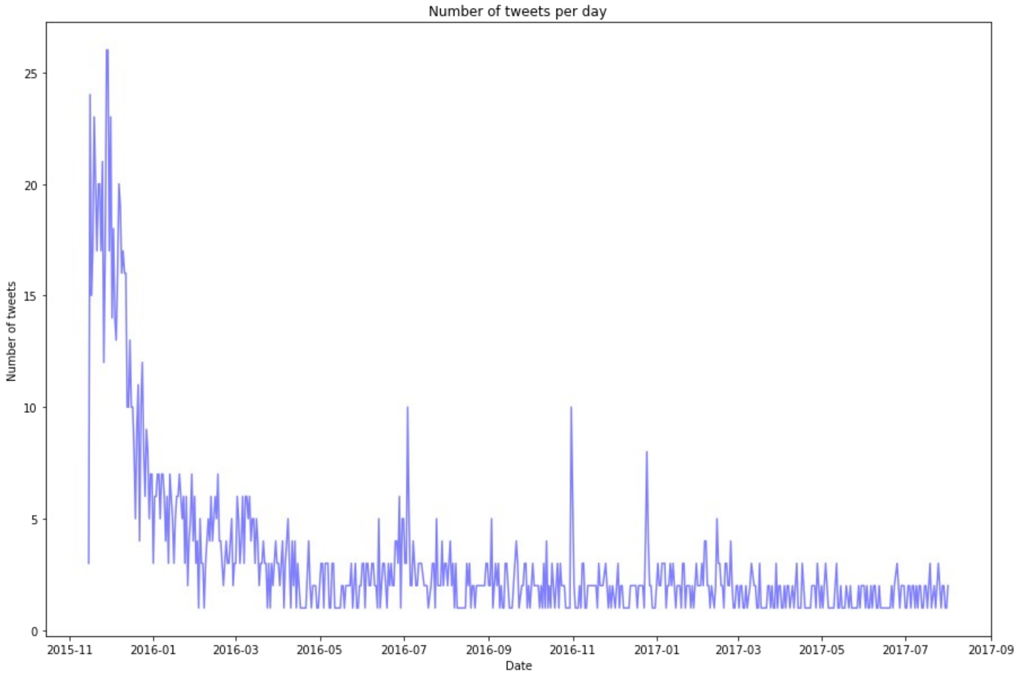
It is also possible to analyze the modification in user interaction over time. In the following plot, it is possible to notice an increasing trend in the number of favorite counts over time, while retweets stay with a more constant number. This could be a possible indication that users prefer to interact with tweets by favoriting them instead of retweeting.

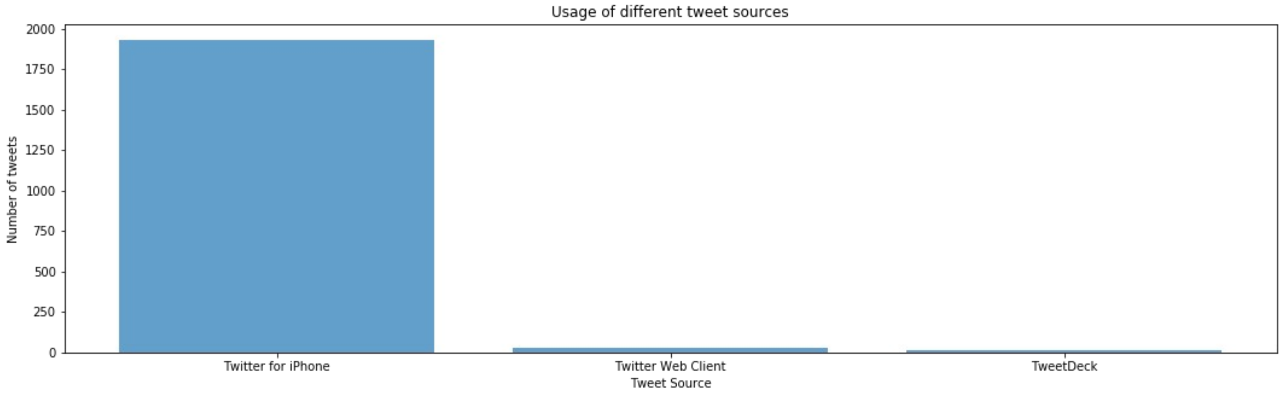


The relation between retweets and favorites can be further analyzed. It is possible to notice an increasing relationship between those variables, indicating that a higher number of retweets could also lead to a higher number of favorites. This can be observed in the following plot:



As a final remark, it is possible to mention that the number of tweets made by WeRateDogs had a peak in late 2015 and decreased, reaching a stable and continuous trend over time. Also, the most preferred source for tweeting was the iPhone. Both this assumptions can be observed in the following plots.





## Conclusions

Regarding the final DataFrame composed by the data obtained in the wrangling procedure, it is taken as final conclusions:

• Golden Retriever is the most popular dog breed, it was the most mentioned in tweets and it also received higher ratings.

• Oliver is the most popular dog name for this dataset.

• The dataset has most dogs identified as in pupper stage.

• Dog ratings are usually high, with plenty of values above 10.

• There is an increasing trend in the number of favorite counts over time, while retweets stay with a more constant number. This could be a possible indication that users prefer to interact with tweets by favoriting them instead of retweeting.

• There is a strong increasing relation between retweets and favorites.

• The number of tweets made by WeRateDogs had a peak, but now it has a constant number of tweets per day.