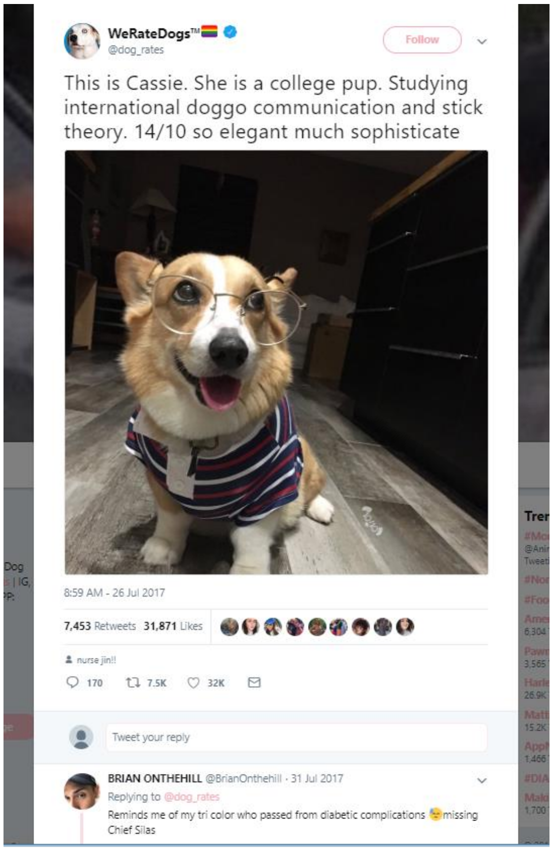
Insights and Visualisations of analysis performed on tweet archive of Twitter user @dog_rates

Introduction:

The dataset that was wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. 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. But the numerators are almost always greater than 10. 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage. Using Python and its libraries, tweet data has been gathered from a variety of sources and in a variety of formats, assessed for its quality and tidiness, then cleaned. This is called data wrangling. I have documented the insights and visualisations produced from the wrangled data using Python in this notebook (and its libraries).



WeRateDogs™ @dog_rates

This is Cassie. She is a college pup. Studying international doggo communication and stick theory. 14/10 so elegant much sophisticate

8:59 AM - 26 Jul 2017

7,453 Retweets 31,871 Likes

nurse jin!!

170 7.5K 32K

BRIAN ONTHEHILL @BrianOnthehill · 31 Jul 2017

Replying to @dog_rates

Reminds me of my tri color who passed from diabetic complications 😓missing Chief Silas

An example tweet 1

Insights:

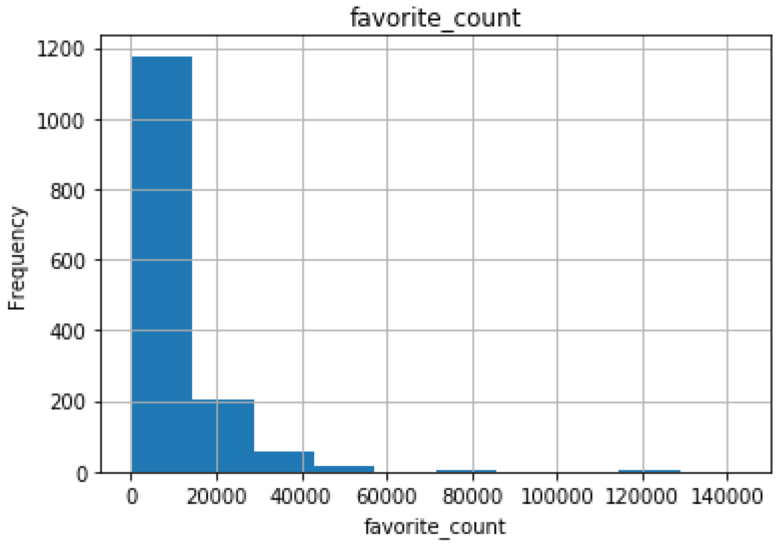
- Here are the top 10 most popular breeds of dogs among pet owners.

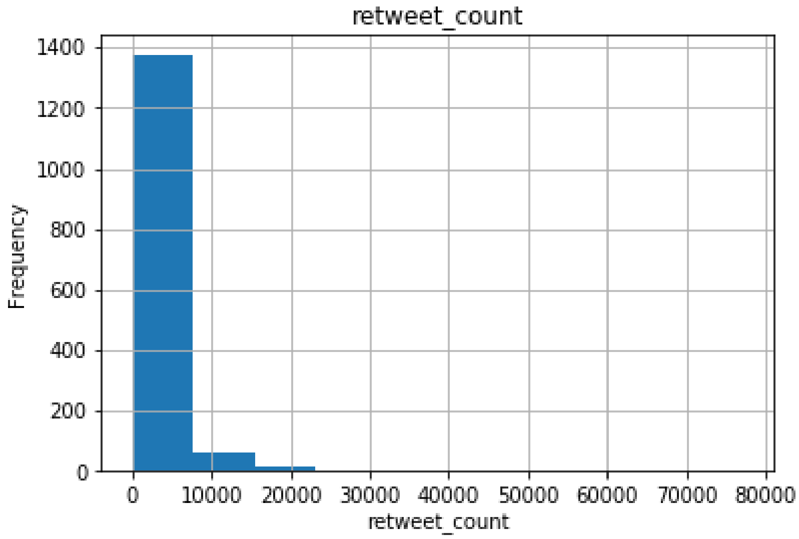
1. golden_retriever
2. Labrador_retriever
3. Pembroke
4. Chihuahua
5. pug
6. chow
7. Samoyed
8. Pomeranian
9. toy_poodle
10. Malamute

- Though, most of the tweet ratings have a denominatior 10, tweets with many dogs in the image are given ratings with higher denominators.
- Tweets with interesting or emotional or funny stories/videos about the dog seem to result in higher retweet and favorite counts.
- Tweets with normal looking and no special features seem to be given medium ratings, with ratings ranging from 2 to 14 for denominator 10.
- On the other hand, tweets that use negative words to describe the dog or posts about anything other than a dog or have awkward/not very good looking pictures are given lowest ratings.

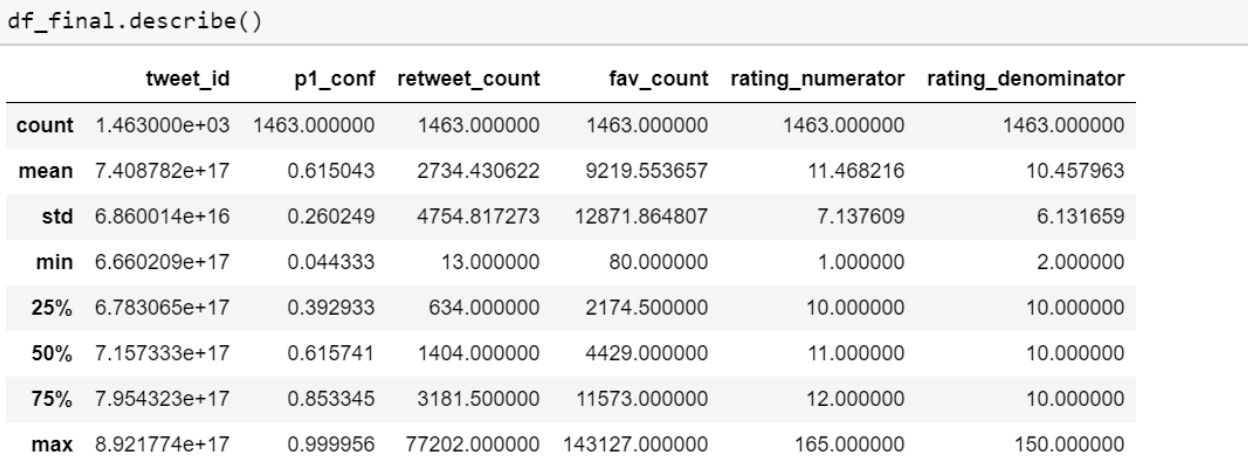
Golden Retriever 1

Visualisations:



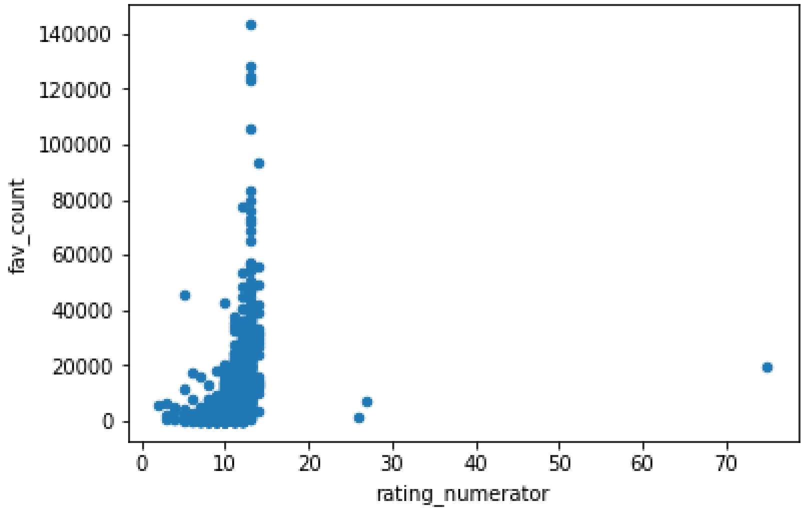


The above images are histograms of retweet_count and fav_count. It can be observed that most of the values of retweet_count are below 8000 and fav_count below 15000. The screenshot of descriptive statistics table below gives a clearer picture of retweet_count and fav_count numbers.

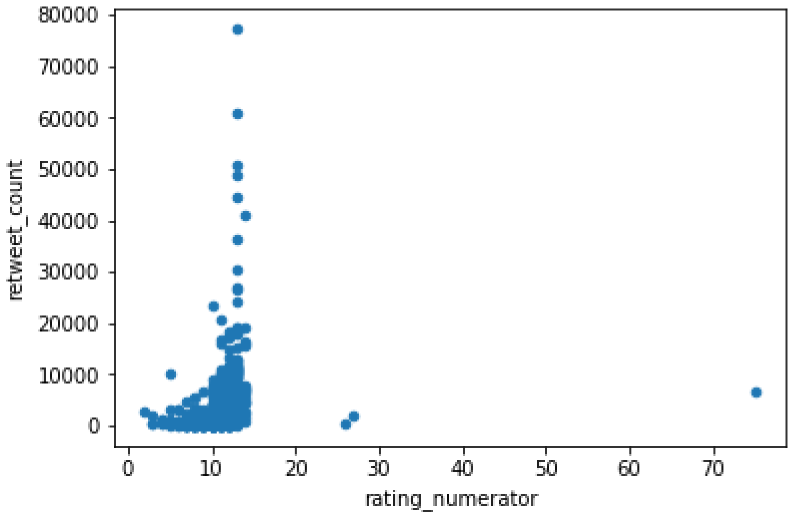


```
df_final.describe()
```

	tweet_id	p1_conf	retweet_count	fav_count	rating_numerator	rating_denominator
count	1.463000e+03	1463.000000	1463.000000	1463.000000	1463.000000	1463.000000
mean	7.408782e+17	0.615043	2734.430622	9219.553657	11.468216	10.457963
std	6.860014e+16	0.260249	4754.817273	12871.864807	7.137609	6.131659
min	6.660209e+17	0.044333	13.000000	80.000000	1.000000	2.000000
25%	6.783065e+17	0.392933	634.000000	2174.500000	10.000000	10.000000
50%	7.157333e+17	0.615741	1404.000000	4429.000000	11.000000	10.000000
75%	7.954323e+17	0.853345	3181.500000	11573.000000	12.000000	10.000000
max	8.921774e+17	0.999956	77202.000000	143127.000000	165.000000	150.000000



Analysis for above plot:
Most ratings are in the range of 2-14. In this plot, an exponential increase in fav_count with increase in rating_numerator can be observed with the curve ending at rating_numerator value of 14. The outliers belong to tweets with images of multiple dogs.



Analysis for above plot:

Retweet count vs rating_numerator plot follows the same pattern as fav_count vs rating_numerator. But this plot is not as dense as the previous plot, especially at higher counts(counts above 20,000) on y axis. This might be because, when we come across interesting posts/tweets, many of us hit the favorite icon rather than retweet.