Analyzing Wrangled WeRateDogs Twitter Data

Chengyu Hang

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. It was started in 2015 by college student Matt Nelson, and has received international media coverage both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter. [1]

In this report, I provide the analysis of the clean WeRateDogs data with 1908 observations. I mainly analyzed;
- the relationships between ratings, favorite counts and retweet counts,
- the average rating and the average favorite counts of the most tweeted dog breeds,
- the average rating and the average favorite counts of the dogs based on their stages.

Analysis

1.The dog with the highest rating

Atticus has the highest rating 1776/10 (177.6). His dog stage and breed are unknown. Observe that his retweet count is very close to the mean retweet count and its number of likes is much more lower than the mean favorite count which is 8489.4. Atticus dressed with the US flag on the indepence day of America, 4th of July, so it explains why this rating is so high.

Note: I changed his rating to 1.5 because it would be difficult to work with such a high number in the further data analysis. It will still be the highest rating.

2. Favorite Dog

This doggo without a name has the highest number of likes and it is a Labrador Retriever. He also has the highest retweet count. While he is swimming he suddenly realizes that he can stand in the pool, it is hilarious.

3. Descriptive Statistics

	retweet_count	favorite_count	rating	confidence
count	1908.000000	1908.000000	1908.000000	1908.000000
mean	2591.144130	8489.412998	1.061204	0.466313
std	4703.502207	12634.236572	0.210510	0.339239
min	12.000000	78.000000	0.200000	0.000000
25%	577.750000	1773.750000	1.000000	0.145375
50%	1245.000000	3777.500000	1.100000	0.457514
75%	2949.500000	10608.250000	1.200000	0.778514
max	83626.000000	164232.000000	1.500000	0.999956

This table will guide us in the data analysis.

4. Distribution of the Ratings

Distribution of the Ratings
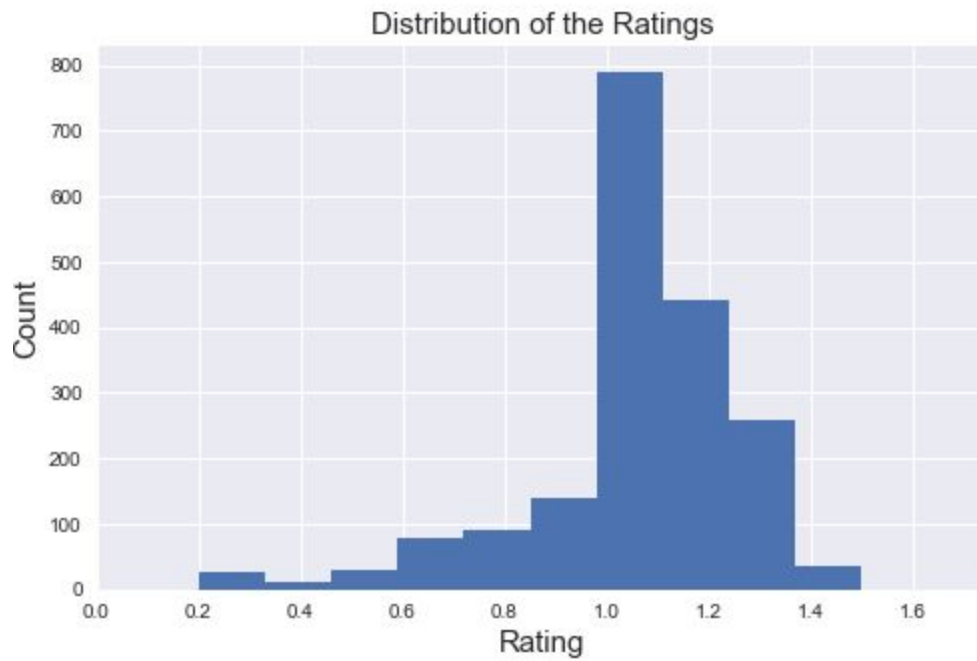
- The mean dog rating is 1.06 and the ratings are more frequent between 1 and 1.3, i.e between 10/10 and 13/10.
- 1.2, 12/10, is the most frequent rating.
- The median value for rating is 1.1, 11/10, which means that half of the ratings are greater than or equal to 1.1.
- Most of the low ratings are not actually dog ratings but I could not clean all of them because it requires too much effort.

5. Relationship between Favorite Count, Retweet Count and Rating

One might think that retweets and favorites are positively correlated and retweets and favorites increase with ratings. So I analyzed the relationship between these variables.

Number of Retweets and Favorites based on Ratings

Distribution of Retweets and Favorites based on Ratings
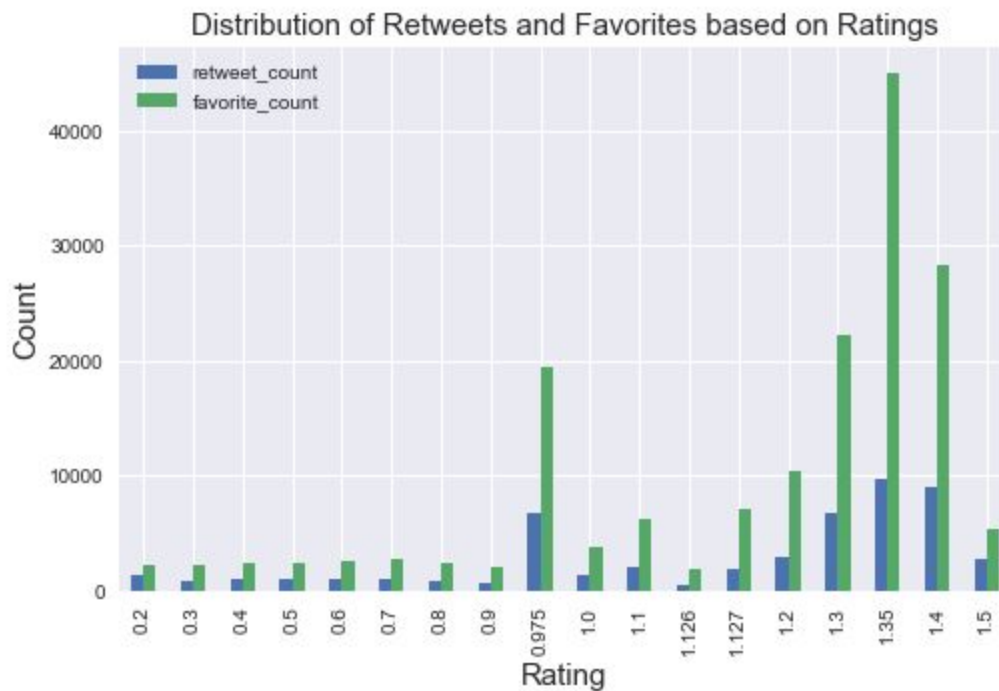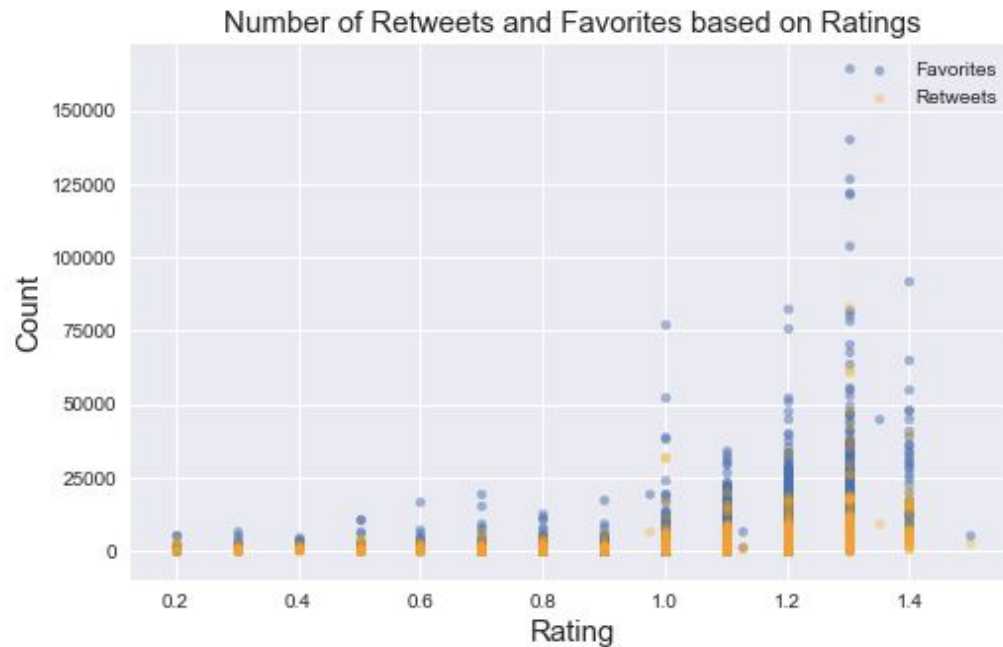
- The favorite counts are higher than the retweet counts for each rating.

- The dogs with the ratings between 1.3 and 1.4 are retweeted and liked most.

It seems like there is a positive correlation between retweet counts and favorite counts based on ratings.

Let's look at the Pearson correlation coefficients:

	rating	retweet_count	favorite_count
rating	1.000000	0.306295	0.406224
retweet_count	0.306295	1.000000	0.930268
favorite_count	0.406224	0.930268	1.000000

- They show that there is indeed a positive correlation between retweet counts and favorite counts and this coefficient is really high, 0.93, which is plausible because people tend to retweet the tweets that they like.
- Surprisingly there is no correlation between ratings and retweet counts, as well as ratings and favorite counts. One possible reason is that those ratings are given by only one account holder WeRateDogs and they are quite subjective. It is normal other people to like or retweet the tweets independent from their ratings but considering the pictures. Another reason might be the harmony between the picture and the tweet, if the text or the image are considered funny by the followers it might get higher number of likes and retweets even if the dog has a low rating.
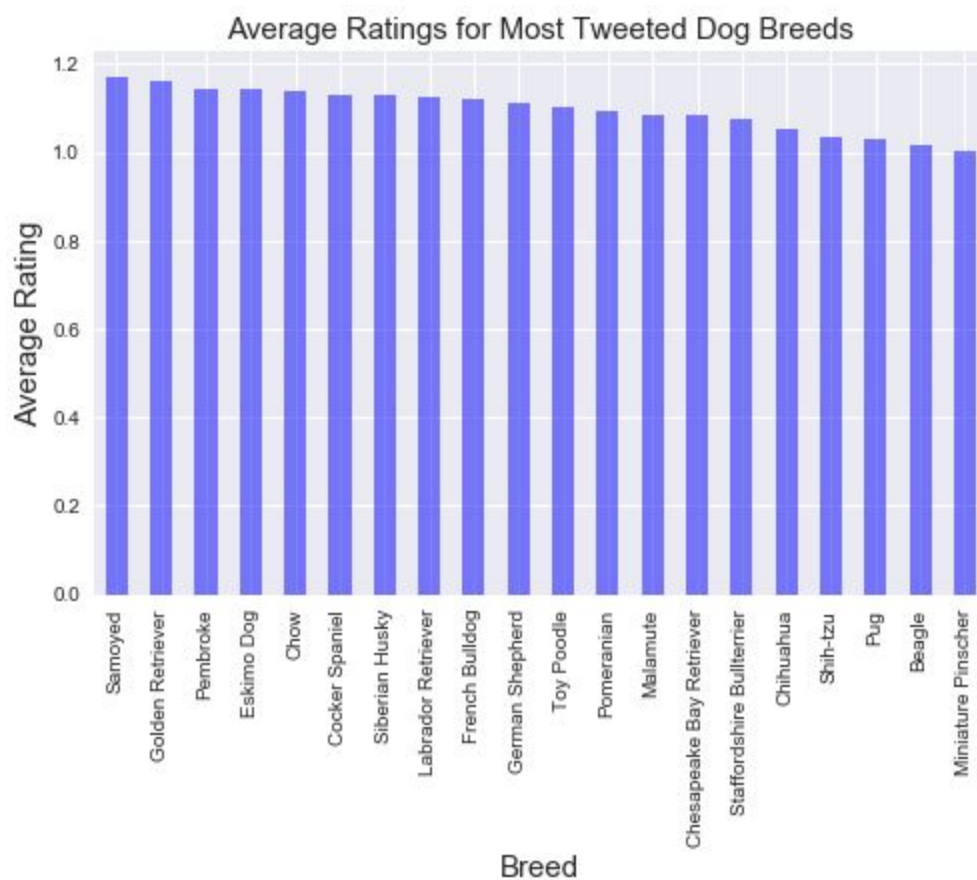
6. Analysis of Most Tweeted Dog Breeds
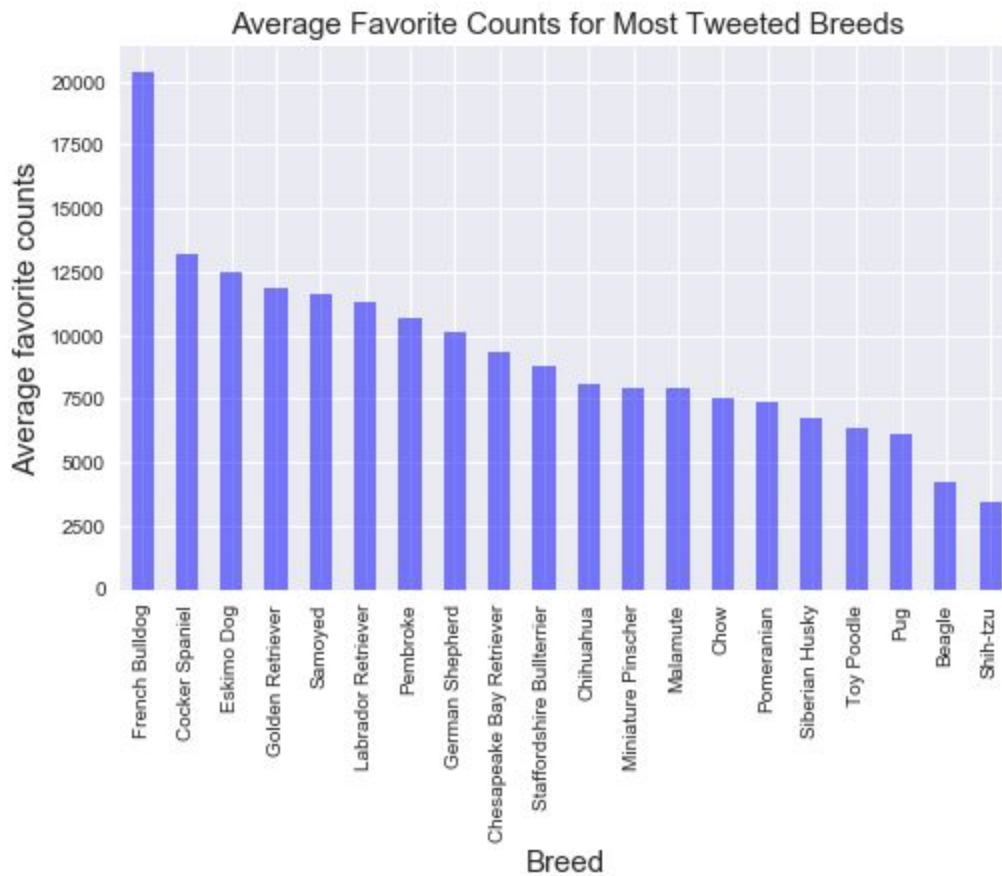
Most Tweeted Dog Breeds
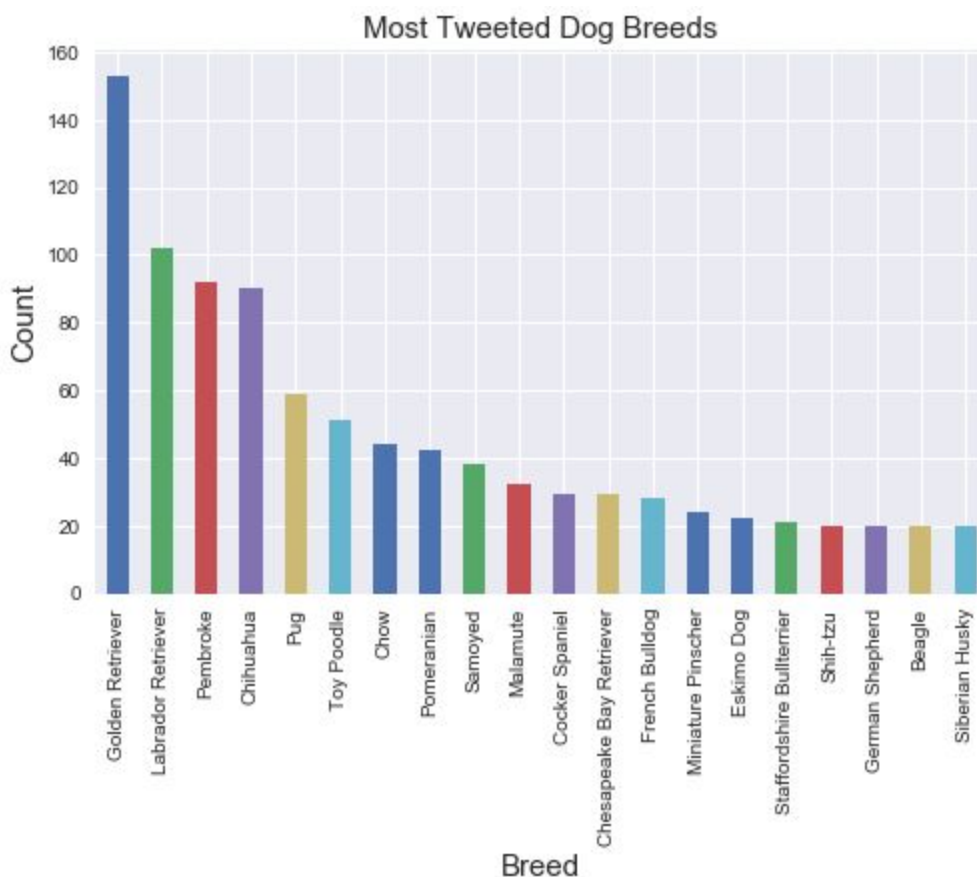
I only took into account the breeds with at least 20 tweets and the most tweeted dog breeds in order: Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, Pug. The number of Golden Retriever is substantially higher.

Average Ratings for Most Tweeted Dog Breeds

Most tweeted dog breeds with the highest rating in order: Samoyed, Golden Retriever, Pembroke, Eskimo Dog, Chow.

Average Favorite Counts for Most Tweeted Breeds

Average Favorite Counts for Most Tweeted Breeds

- Most tweeted dog breeds with the highest number of likes: French Bulldog, Cocker Spaniel, Eskimo Dog, Golden Retriever, Samoyed.
- It looks like **French Bulldog** is the most favorite dog breed.

Most liked breeds differ from the breeds with highest ratings. **Samoyed**, **Golden Retriever**, and **Eskimo Dog** are common in the both lists.
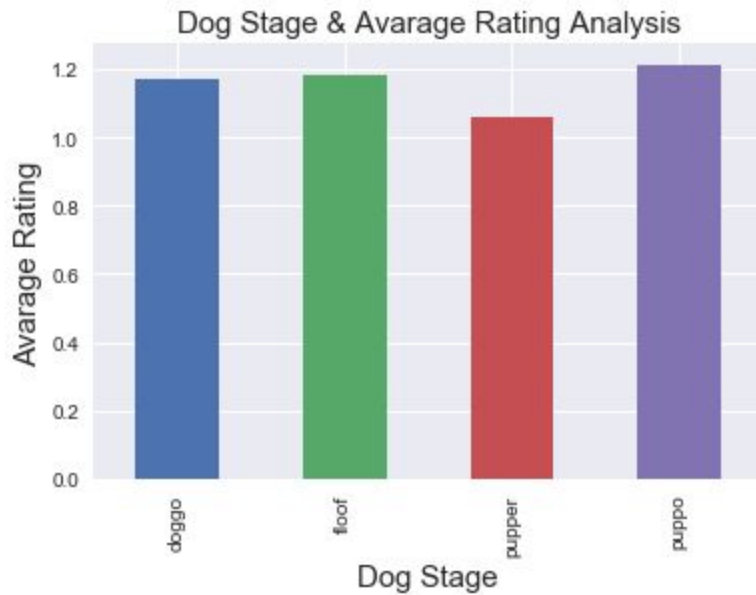
7. Dog Stage Analysis

Dog Stage Distribution

There are 1566 dogs with unknown dog stage. It is a very high number and more time should be spent on cleaning dog_stage part. However, for the current analysis, I excluded the unknown dog stages.
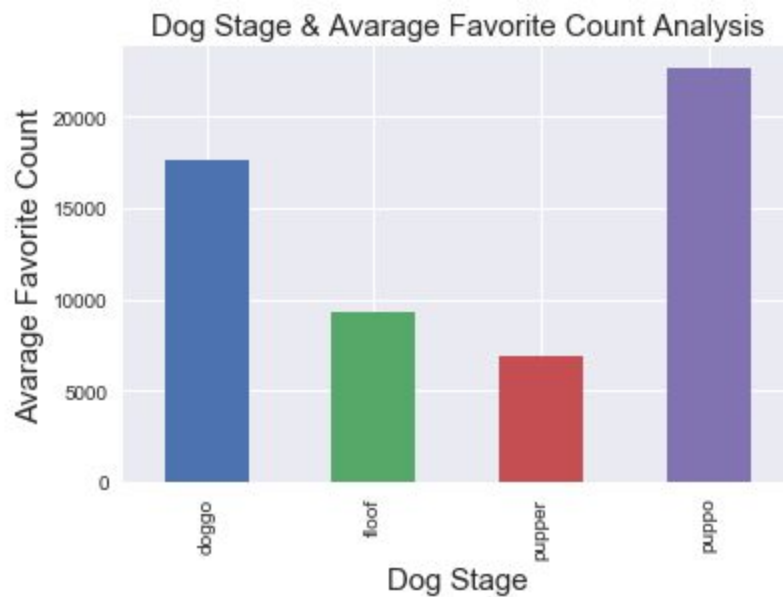

Dog Stage Distribution
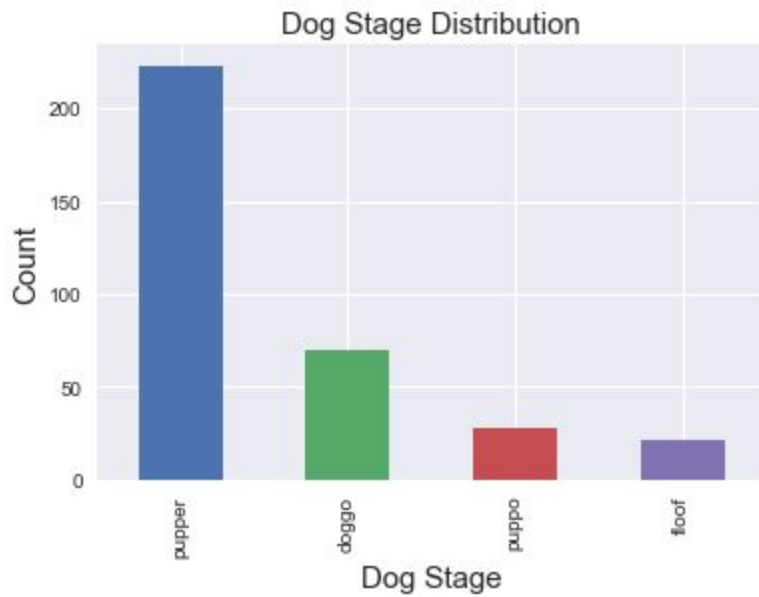
Most of the dogs are pupper.

Dog Stage & Average Rating Analysis

The highest average rating is owned by puppos. Since it is an account for fun, it makes sense that all of them have similar and good ratings.

Dog Stage & Average Favorite Count Analysis

- The puppos has the highest favorite counts.
- Although, the puppers are more tweeted about, they have the lowest rating and favorite counts.

- As I stated before it is not surprising to have such differences between ratings and like counts because the ratings are given by a unique account holder WeRateDogs in a subjective way.

Remarks:

- The ratings, retweets and favorites are not only related with the dog images or videos, the humor of the account holder also affects the followers' decisions.
- Image prediction algorithm is not always successful to detect the dog breed. Sometimes it detects another object in the picture rather than the dog, like a bagel.
- I believe I did a good job in re-extracting, cleaning and correcting names and ratings but dog stages needs much more work rather than extraction. Most of the dogs, their stages are not stated in the tweets.
- Some tweets rate other things rather than dogs, I cleaned a significant amount of data but it was not that easy to clean all of them, that is why there are still very low ratings.