Analysis of a Dataset Downloaded from WeRateDogs Archive

From Wikipedia:

> "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.
>
> WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10". "

Description of the Data

The data used to create this analysis comes from three different sources:

- An enhanced Twitter archive provided by Udacity. This dataset contains basic tweet data for more than 2000 tweets. The tweet text was used to extract ratings, dog names and dog stages (these are, in the WeRateDogs' language 'doggo', 'puppo', 'pupper' and 'floofer').
- Additional data obtained by querying the Twitter API; this contains each tweet's retweet count and favorite ("like") count.
- Results obtained by running a neural network on the Twitter archive in order to predict what breed of dog (or other object, animal, etc.) is present in each tweet.

The three dataframes are merged into a single Pandas dataframe. The data is wrangled and the cleaned tidy dataset contains 1993 entries and 22 columns.

Data Analysis

Unusual ratings. Unusual findings.

WeRateDogs ratings are given in fractional form, with denominator of 10 (in most of the cases) and a numerator that is usually an integer between 10 and 14. However, other ratings can be found. I decided to take a closer look at some of these unusual ratings.

To start with consider those ratings where the denominator is not 10. There are 12 such ratings in our dataset and they are listed below:

$$\frac{84}{70},\ \frac{165}{150},\ \frac{204}{170},\ \frac{99}{90},\ \frac{80}{80},\ \frac{45}{50},\ \frac{60}{50},\ \frac{44}{40},\ \frac{143}{130},\ \frac{121}{110},\ \frac{144}{120},\ \frac{88}{80}$$

In each rating the denominator is a multiple of 10. It turns out that all but one denominator is of the form $10 \cdot n$ where n is the number of dogs in the tweet's image. The tweet that does not follow this pattern is the one with rating $\frac{143}{130}$.

To continue this analysis, notice that the remaining 11 numerators are in fact multiples of the same n. Simplify each fraction by n to get:

$$\frac{12}{10},\ \frac{11}{10},\ \frac{12}{10},\ \frac{11}{10},\ \frac{10}{10},\ \frac{9}{10},\ \frac{12}{10},\ \frac{11}{10},\ \frac{143}{130},\ \frac{11}{10},\ \frac{12}{10},\ \frac{11}{10}$$

Thus these ratings fall in the usual range of values. In an amusing way, each rating represents the cumulative ratings of the n dogs in the tweet (assuming that all dogs in the tweet get equal ratings, of course).

For example, the image from the tweet with rating $\frac{204}{170}$, contains 17 dogs:

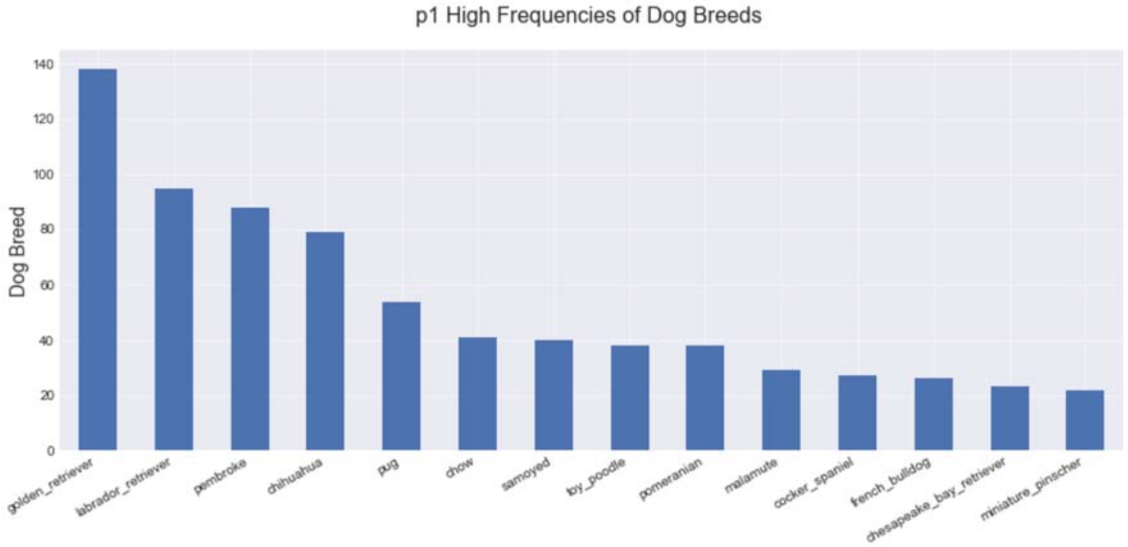
Next, consider the unusually large numerator values. Among the 20 largest values, only 14 are greater than 14 of which only 2 values correspond to tweets not included in our previous analysis of multiple dogs tweets. These two ratings are $\frac{1776}{10}$ for a dog wearing a 4th of July outfit

and $\frac{420}{10}$ for a picture of Snoop Dog. Meaningful funny ratings!

Dog breeds analysis

The dog breeds predictions of a neural network are given in columns 'p1', 'p2' and 'p3'. Confidence levels for each prediction are also available. Since the mean confidence level for prediction p1 is 0.59 and the next mean confidence level for prediction p2 is about 0.13, I will analyze p1 only.

The frequency histogram for the dog breeds that appear at least 20 times in prediction p1 indicates that golden retriever is the most tweeted dog breed.



It turns out that there are 138 occurences of golden retriever in p1, two of which have unusual ratings of $\frac{99}{90}$ and $\frac{143}{130}$ (see the above paragraph). I consider those tweets outliers and not include them in the following analysis. The basic statistics data for the 136 predicted golden retrievers is given below:

golden retriever statistics

```
       rating_numerator       retweet_count   favorite_count     p1_conf
count        136.00                 136.00           136.00       136.00
mean          11.63                3613.05         12315.39         0.72
std            1.20                4361.46         13024.53         0.22
min            8.00                  51.00           192.00         0.14
25%           11.00                1175.25          3520.25         0.60
50%           12.00                2244.00          8134.50         0.77
75%           12.00                4250.25         16075.75         0.90
max           14.00               26972.00         83573.00         0.99
```

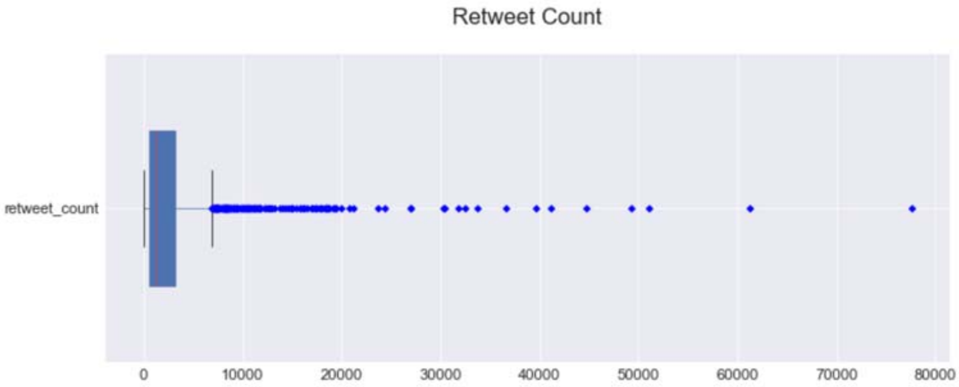
The smallest rating numerator is 8 (all rating denominators are 10 for this dataset) while the maximum is 14. The highest favorite count is over $80,000$ which, according to Matt Nelson, shows that the post went viral. The confidence level for identifying the golden retriever is quite high, with an average of 72%, which indicates that the neural network has better chance (compared with 59% overall confidence) of identifying this breed. This might be another reason for seeing a higher frequency for golden retriever in comparison with other dog breeds.

Retweet and favorite count analysis

Let's take a look now at statistics for the 'retweet_count' and 'favorite_count' columns. To better understand the distributions I also draw the two boxplots.

retweet_count statistics

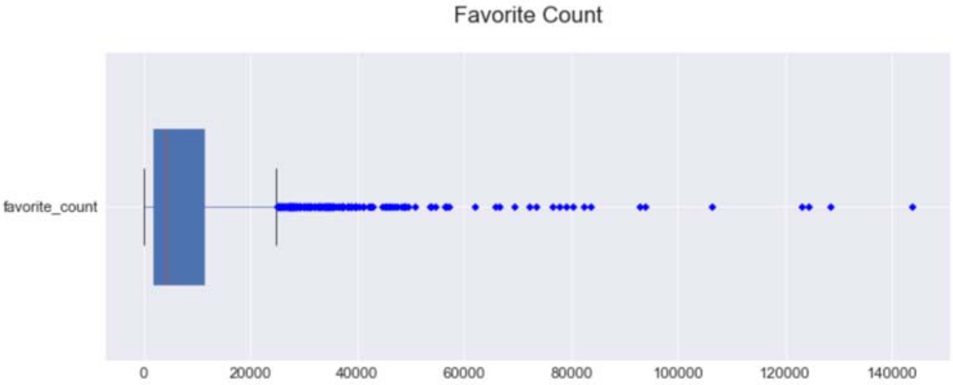
```
count     1993.00
mean      2725.33
std       4705.58
min         13.00
25%        609.00
50%       1312.00
75%       3132.00
max      77544.00
```



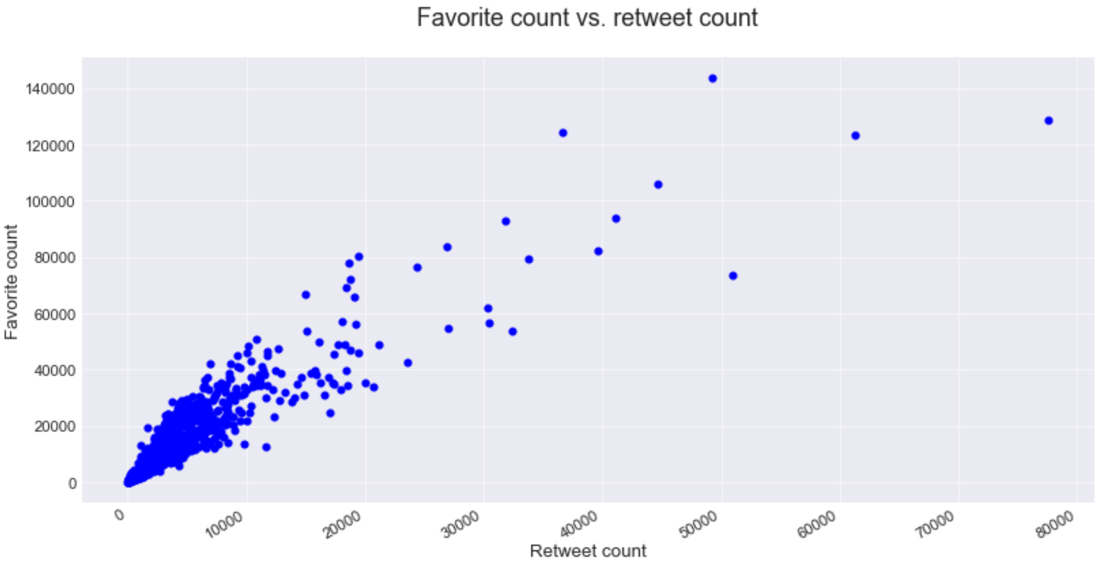
The statistics and the boxplot for the favorite count are:

favorite_count statistics

```
count      1993.00
mean       8871.58
std       12595.22
min          79.00
25%        1926.00
50%        4050.00
75%       11173.00
max      143702.00
```



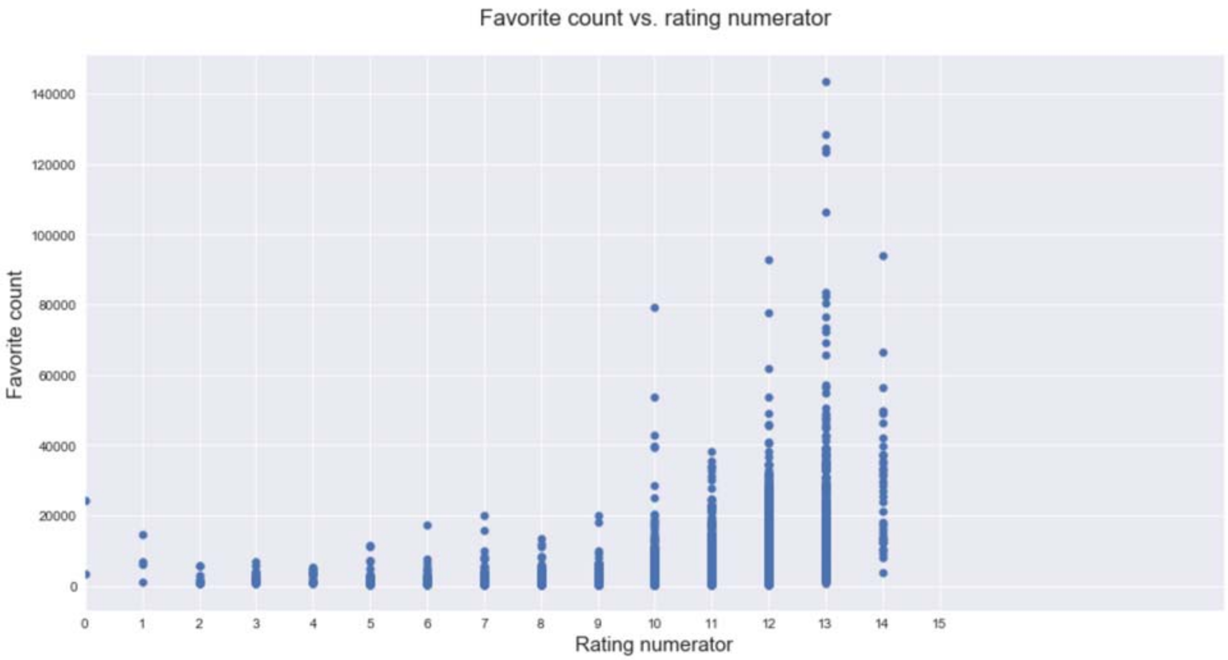
The mean number of retweets is less than 3000, while for the favorite counts is almost 9000. In both cases there are numerous outliers, among which several are extreme. Are these outliers related? With other words, are the tweets with the most retweets the same with the ones that get most favorite counts? To see this we plot these parameter sets together:



Clearly more a post is retweeted, more 'likes' (favorite_count) will gets. However, the tweet with most favorite counts is not the same with the one that gets most retweets.

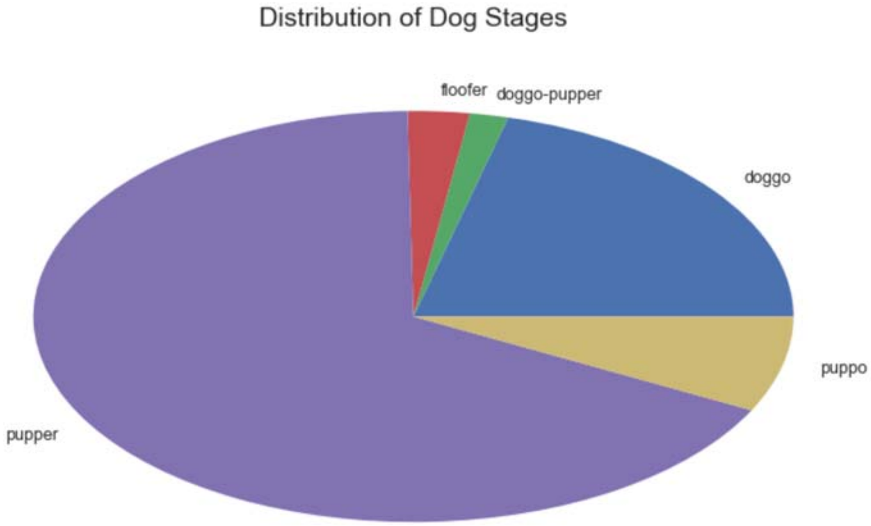
As a fun fact, the post with most favorite counts is of a dog marching in the 2017 Women's March, which was retweeted more than $50,000$ times and favorited more than $143,000$ times (see the Wikipedia page for WeRateDogs).

From the following plot, it is clear that there is also a direct relationship between the ratings and the favorite coun:



About dog stages

WeRateDogs created an internet language for people who love dogs, for example the dog stages from our dataset. To conclude with, I take a look at the frequency chart of the dog stages. It turns out that the puppers are the most popular doggos out there.



Conclusion

The data extracted from the Twitter archive WeRateDogs is quite interesting and fun to work with. Some aspects were discussed here, but there are many more facets of this data that can be analyzed, they could involve tweet time and date, the contents of the tweet text or even the popularity of certain dog names. The ratings system is quite original and the site helped create a popular unique dog jargon.