Analysis and Insight of the WeRateDogs Data

The analysis is done after assessment and cleaning has been carried out on the datasets obtained from the weratedog. The data was saved as `twitter_archive_master.csv` file and read into the notebook as a dataframe table named `master_table`.

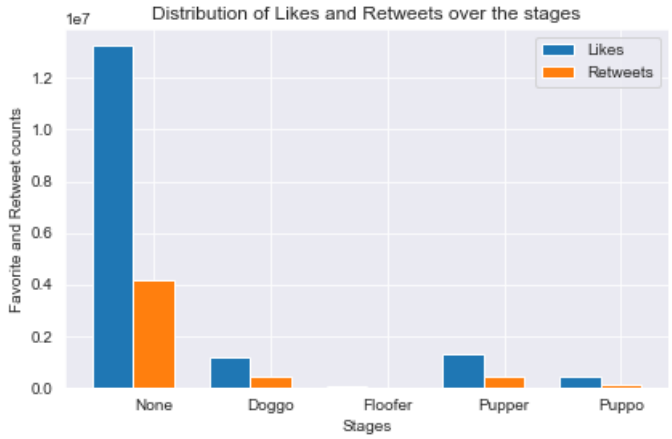
The analyses were carried out to answer the following questions:

1. What are the percentages of retweet and favorite counts for the dog stages?

The data in the master_table is grouped by dog stages using the groupby method and agg() function used to get the sum of favorite count and retweet count for each of the dog stage. This data is saved in a variable called stages_percent. Then lambda functions are used to compute the favorite and retweet count percentages and `% likes` and `% retweets` columns are added to the table to show the percentages.

Here is the table:

stages	favorite_count	retweet_count	% likes	% retweets
None	13229109	4188963	81.171617	79.432190
doggo	1164400	463841	7.144565	8.795472
floofer	99659	33328	0.611491	0.631974
pupper	1346066	455385	8.259238	8.635127
puppo	458469	132117	2.813090	2.505236



The group barchart shows that the favorite count is more than retweet count for every dog stage

2. What are the distributions of retweet and favorite counts over time?

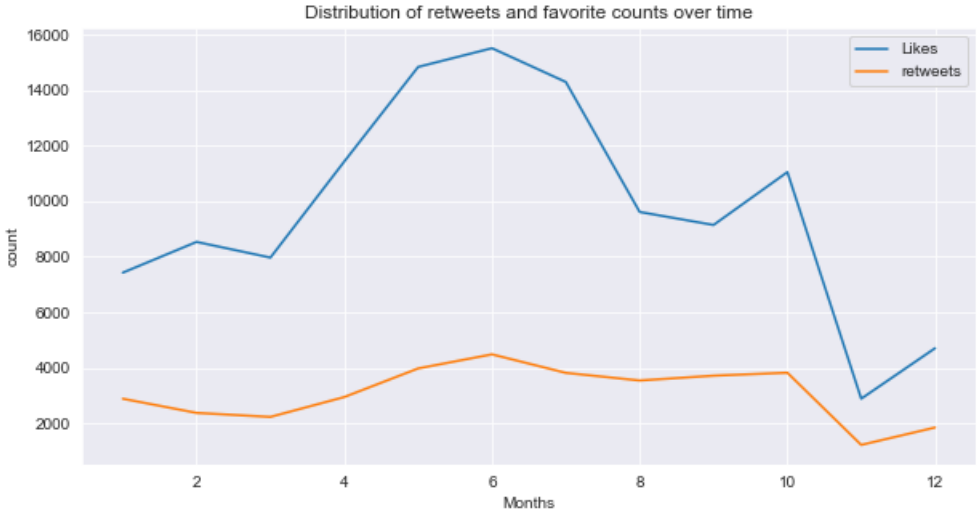
The tweet_id, favorite_count, and retweet_count columns were extracted from the master_table and saved as `time_tweet`. Then, a new column named `month` which store the months extracted from the timestamp column in the master_table using pandas `DatetimeIndex` method, was added and the mean of the favorite and retweet count were computed as I grouped the time_tweet table by months using groupby function.

Here is the table:

month	retweet_count	favorite_count
1	2887.493827	7424.423868
2	2376.534091	8533.147727
3	2232.293413	7967.904192
4	2947.134021	11427.731959
5	3976.950495	14832.722772

6	4484.368000	15507.328000
7	3817.808511	14290.581560
8	3543.000000	9611.200000
9	3716.242857	9142.514286
10	3823.070423	11046.436620
11	1223.956284	2887.295082
12	1850.639151	4706.037736

The table above shows that the month of June had the highest average retweet and favorite count. The line plot of both retweet and favorite count gives a clearer picture of the table.



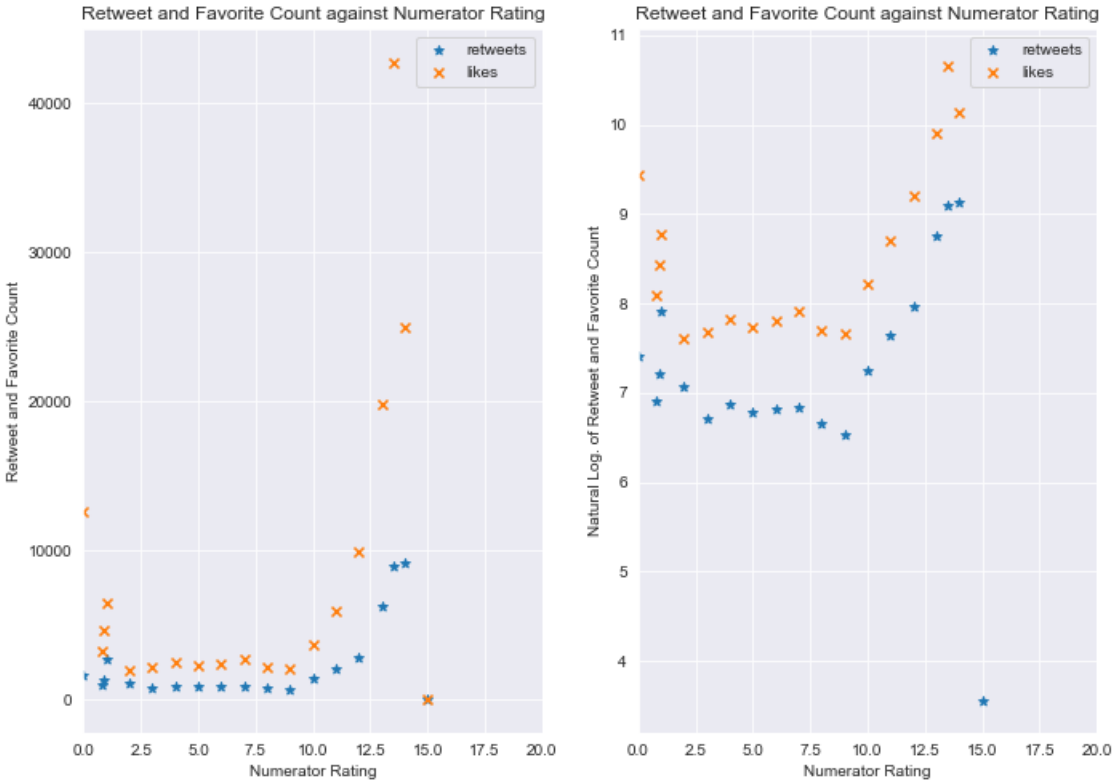
3. How are the numbers of retweet counts and favorite counts distributed over the rating_numerator?

In order to provide answer to this question, the master_table was grouped by rating_numerator using groupby method and agg() method to find the mean of retweet and favorite count for each rating.

rating_numerator	retweet_count	favorite_count
0.0	1648.500000	12614.500000
0.8	994.000000	3267.000000
0.9	1362.666667	4626.666667
1.0	2734.500000	6475.500000
2.0	1165.555556	2023.777778
3.0	825.736842	2149.210526
4.0	958.000000	2476.333333
5.0	885.484848	2294.787879
6.0	906.375000	2451.062500
7.0	926.549020	2728.529412
8.0	782.542553	2202.021277
9.0	690.033784	2127.337838
10.0	1402.827103	3726.724299
11.0	2081.365617	6010.142857
12.0	2866.377919	9961.409766
13.0	6291.775362	19832.898551

13.5	8955.000000	42755.000000
14.0	9173.564103	24988.769231
15.0	35.000000	0.000000
26.0	480.000000	1700.000000
27.0	1628.000000	6596.000000
50.0	205.000000	2346.000000
75.0	6208.000000	18477.000000
420.0	8260.000000	23603.000000
1776.0	2457.000000	5108.000000

The table shows that rating 14 had the highest retweet count and 15 had the lowest, while 13.5 had the highest favorite count and 15 the lowest too. The scatter plots compares the counts.



The first plot is the average retweet and favorite counts against the numerator rating while the second plot is natural logarithm of the retweet and favorite count against the numerator rating. The relationship on the plot is not linear but it looks parabolic and shows a rating of 10 and above having more counts.

In [ ]:

In [ ]: