Analyzing and Visualizing WeRateDogs Data

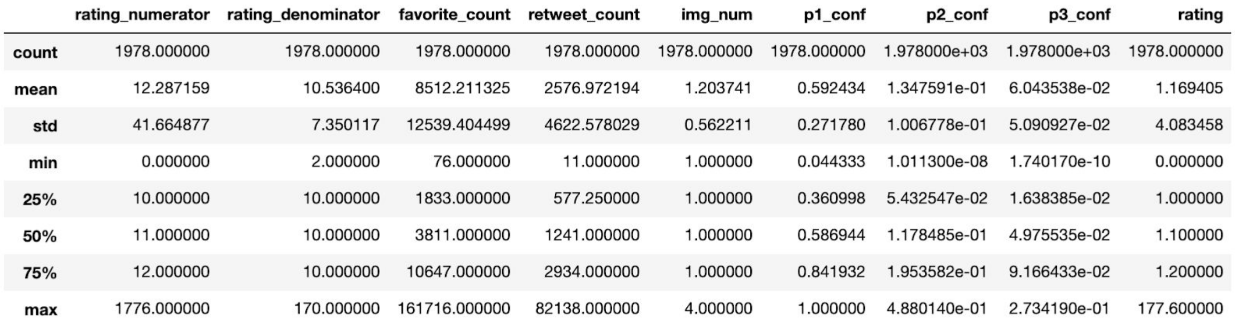
From analyzing and visualizing, I declare some question. The questions are:

1. Are there any outlier in the data?
2. How about correlation between variables?
3. Does the retweet count and favorite count increase with time?
4. Does the rating increase with time?
5. Are the rating affect with the number of favorite and retweet count?
6. How much each algorithm predict the picture is dog?
7. What are the most popular dog names?
8. What are the most popular dog predict?
9. What are the most popular dog predict when all algorithm predict the same dog?

To answer that question, I use library *seaborn* and *matplotlib*. The answer from that questions are:

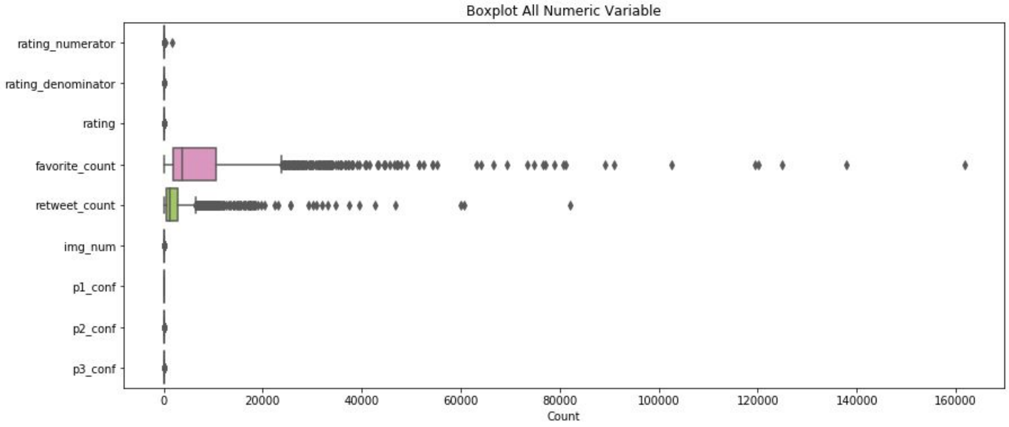
1. Are there any outlier in the data?

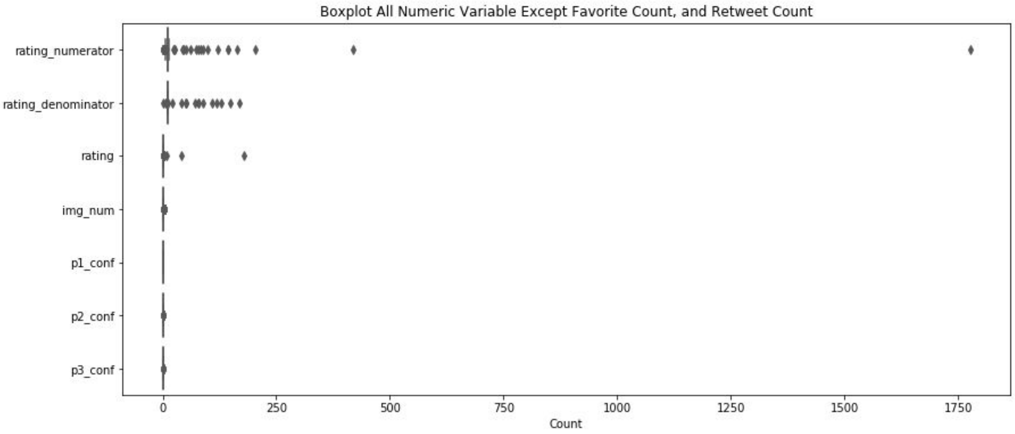
I decide to make stat description and boxplot. From stat description we can see the quantile, mean, and standard deviation. We can know how the data spread.



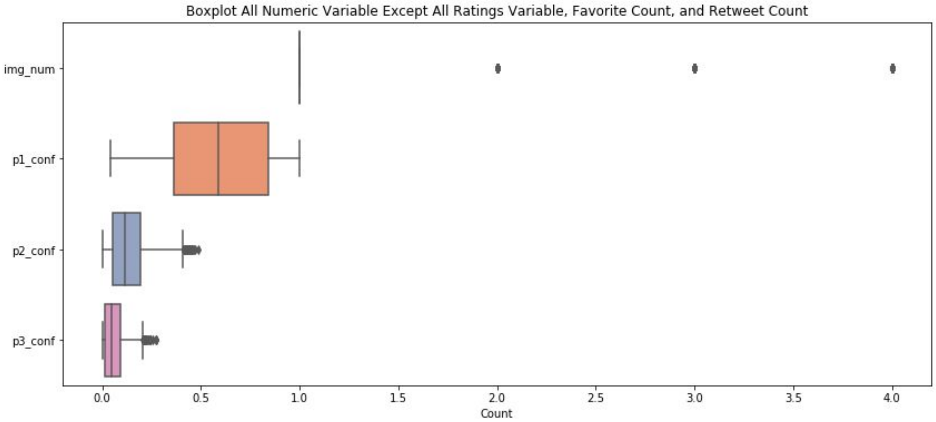
	rating_numerator	rating_denominator	favorite_count	retweet_count	img_num	p1_conf	p2_conf	p3_conf	rating
count	1978.000000	1978.000000	1978.000000	1978.000000	1978.000000	1978.000000	1.978000e+03	1.978000e+03	1978.000000
mean	12.287159	10.536400	8512.211325	2576.972194	1.203741	0.592434	1.347591e-01	6.043538e-02	1.169405
std	41.664877	7.350117	12539.404499	4622.578029	0.562211	0.271780	1.006778e-01	5.090927e-02	4.083458
min	0.000000	2.000000	76.000000	11.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	0.000000
25%	10.000000	10.000000	1833.000000	577.250000	1.000000	0.360998	5.432547e-02	1.638385e-02	1.000000
50%	11.000000	10.000000	3811.000000	1241.000000	1.000000	0.586944	1.178485e-01	4.975535e-02	1.100000
75%	12.000000	10.000000	10647.000000	2934.000000	1.000000	0.841932	1.953582e-01	9.166433e-02	1.200000
max	1776.000000	170.000000	161716.000000	82138.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	177.600000

From the stat desc, we can see that rating_numerator, rating_denominator, and rating have outlier because they have short distance in quantiles, but the max is too far from the Q3. The average, and quantile of img_num is 1, so the another value is outlier. Varibale p1_conf look more normal than p2_conf and p3_conf.

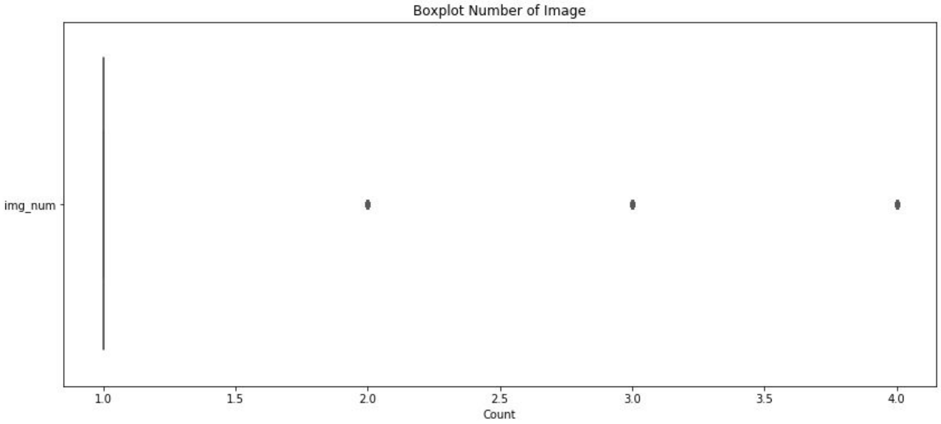




Boxplot All Numeric Variable Except Favorite Count, and Retweet Count



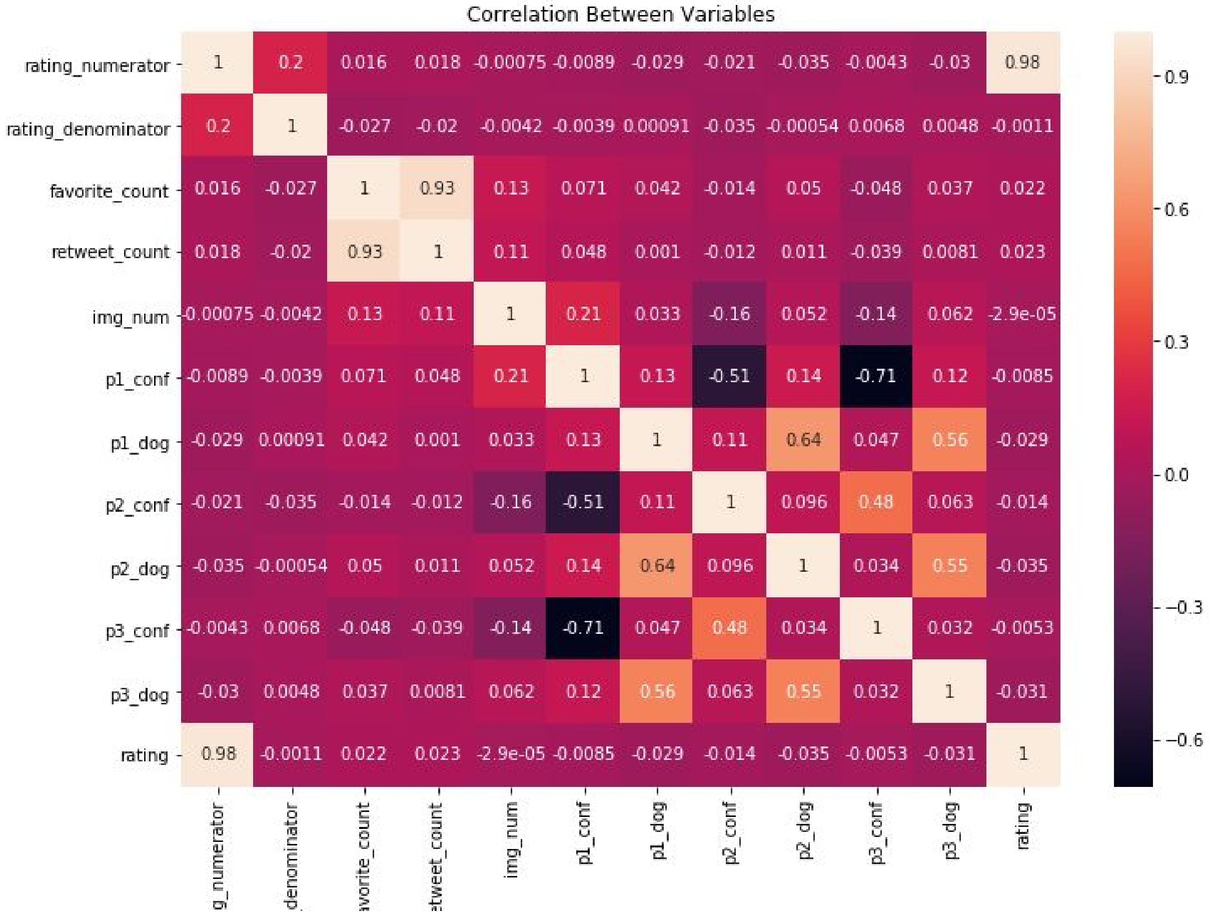
Boxplot All Numeric Variable Except All Ratings Variable, Favorite Count, and Retweet Count



Boxplot Number of Image

From that boxplot we can see that, all numeric cols have outlier except p1_conf. Just like the information from udacity, some nominator have bigger value then their denominator so the rating can be more than 1 (because rating = numerator/denominator so the value must be 0 until 1).

2. How about correlation between variables?



	g_numerator	denominator	avorite_count	etweet_count	img_num	p1_conf	p1_dog	p2_conf	p2_dog	p3_conf	p3_dog	rating
rating_numerator	1	0.2	0.016	0.018	-0.00075	-0.0089	-0.029	-0.021	-0.035	-0.0043	-0.03	0.98
rating_denominator	0.2	1	-0.027	-0.02	-0.0042	-0.0039	0.00091	-0.035	-0.00054	0.0068	0.0048	-0.0011
favorite_count	0.016	-0.027	1	0.93	0.13	0.071	0.042	-0.014	0.05	-0.048	0.037	0.022
retweet_count	0.018	-0.02	0.93	1	0.11	0.048	0.001	-0.012	0.011	-0.039	0.0081	0.023
img_num	-0.00075	-0.0042	0.13	0.11	1	0.21	0.033	-0.16	0.052	-0.14	0.062	-2.9e-05
p1_conf	-0.0089	-0.0039	0.071	0.048	0.21	1	0.13	-0.51	0.14	-0.71	0.12	-0.0085
p1_dog	-0.029	0.00091	0.042	0.001	0.033	0.13	1	0.11	0.64	0.047	0.56	-0.029
p2_conf	-0.021	-0.035	-0.014	-0.012	-0.16	-0.51	0.11	1	0.096	0.48	0.063	-0.014
p2_dog	-0.035	-0.00054	0.05	0.011	0.052	0.14	0.64	0.096	1	0.034	0.55	-0.035
p3_conf	-0.0043	0.0068	-0.048	-0.039	-0.14	-0.71	0.047	0.48	0.034	1	0.032	-0.0053
p3_dog	-0.03	0.0048	0.037	0.0081	0.062	0.12	0.56	0.063	0.55	0.032	1	-0.031
rating	0.98	-0.0011	0.022	0.023	-2.9e-05	-0.0085	-0.029	-0.014	-0.035	-0.0053	-0.031	1

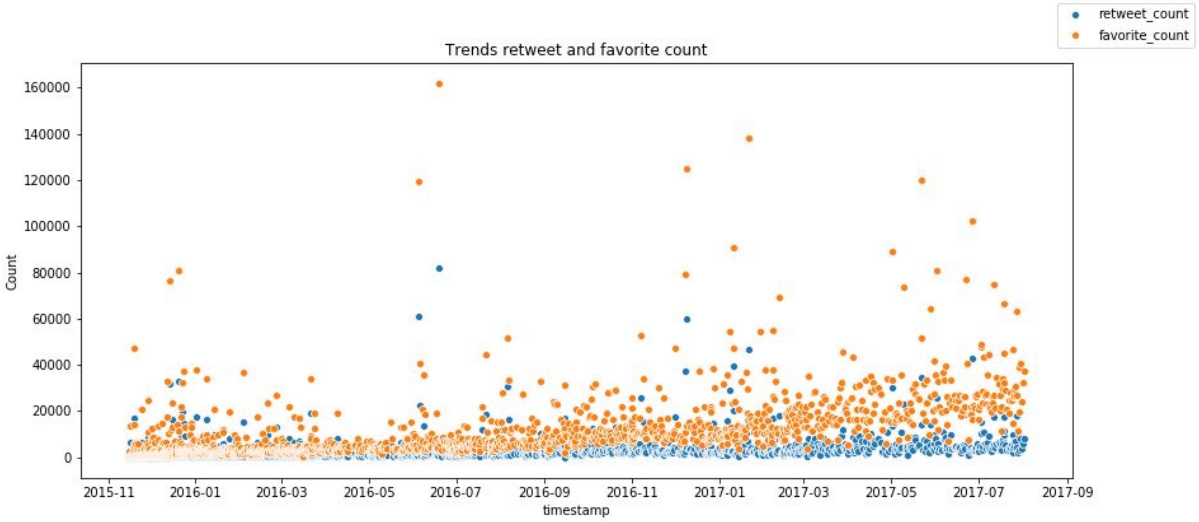
Note: The correlation value is between -1 until 1, negative and positive just to make we know the correlation direction, the closer to the value 0, the smaller the correlation. It use pearson correlation so they just see the linear relationship between each variables.

To this plot please ignore correlation between rating and rating_numerator or rating_denominator because the result should be strong because rating is a calculation from both of them. But surprisingly the correlation between rating and rating_denominator is small. The answer can be found from stat desc that show if value rating_numerator is more varied than rating_denominator (standard deviation of rating_denominator is more hight than rating_numerator but their quantiles just similar each other)

We can see hight positive correlation between favorite_count and retweet_count. Its mean the more favorited the more retweeted and vice versa.

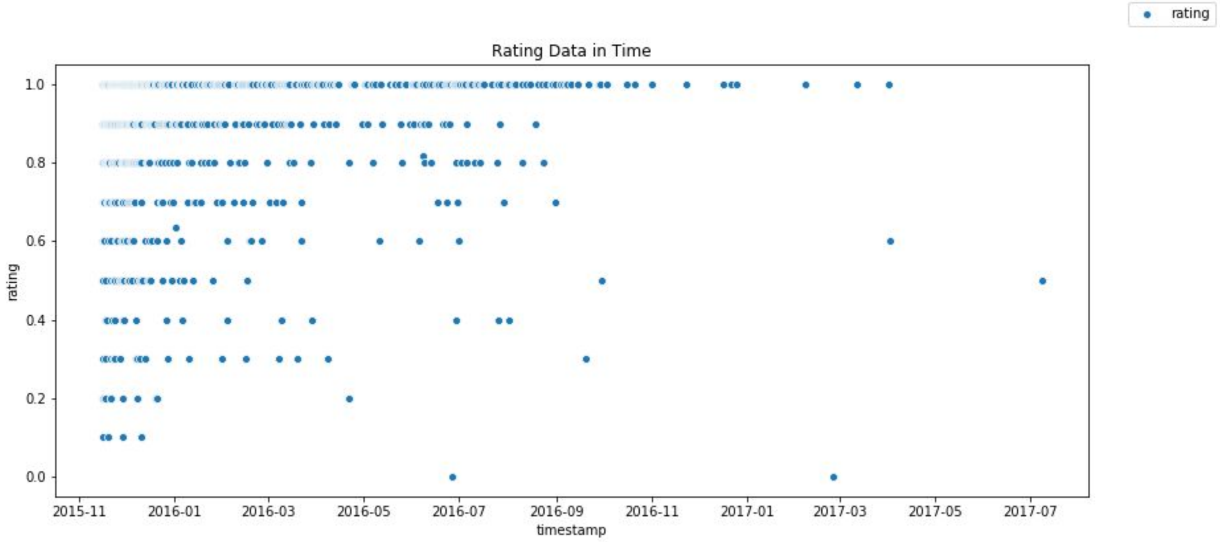
The correlation between all confidence variables also quite high. Somehow when p1_conf hight the confidence in p2 and p3 will decreese, but when confidence p3 increase the confidence in p2 will lightly increese.

3. Does the retweet count and favorite count increase with time?



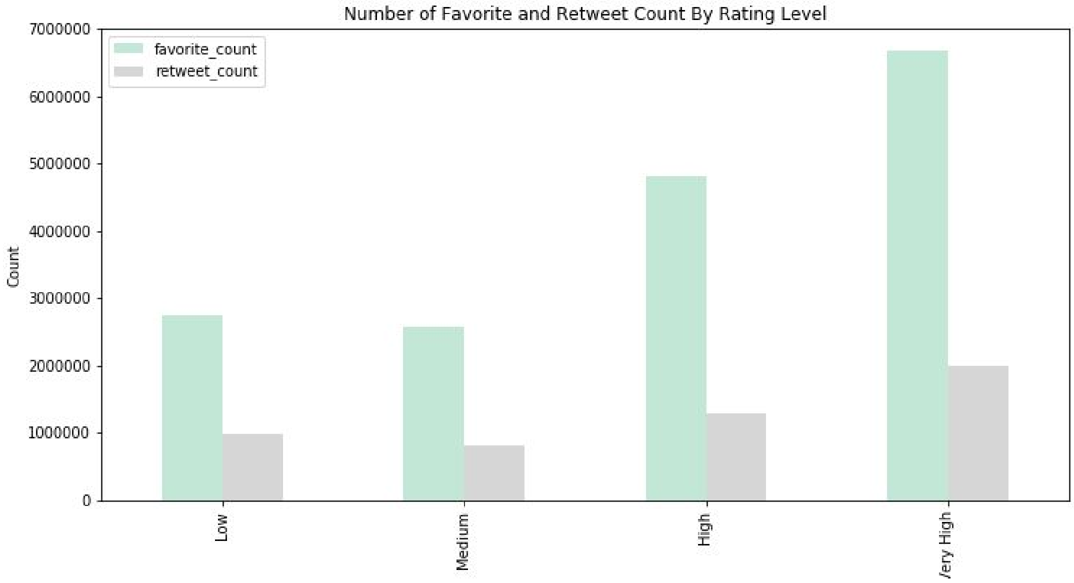
From that visualization, favorite count and retweet count always increase with time. The trends are increasing for both variables. But favorite count growing larger than retweet count

4. Does the rating increase with time?



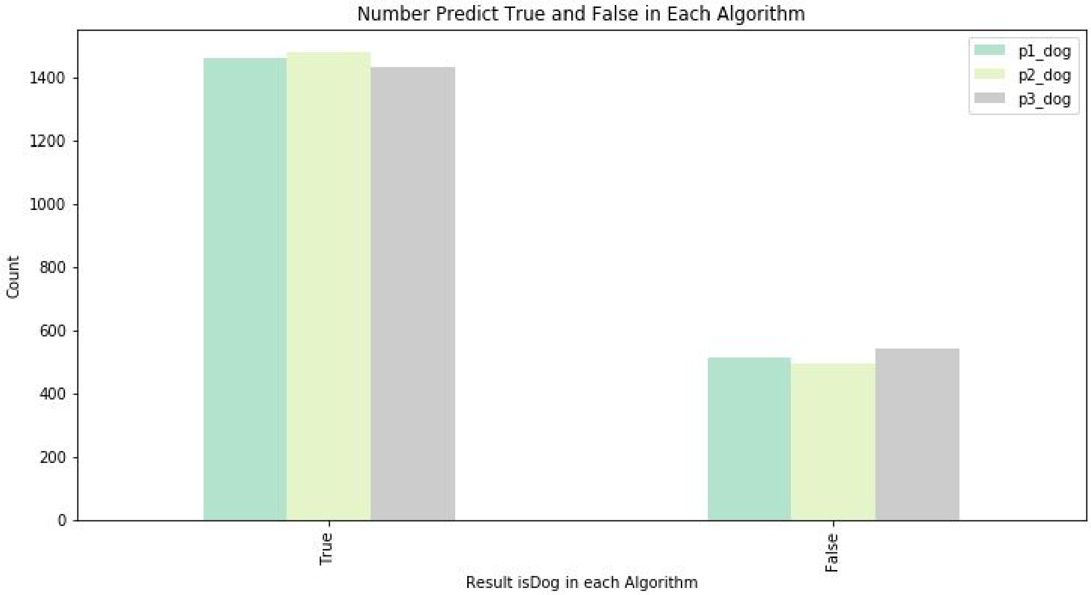
Rating are separated from min to max value at anytime, but from that plot from the same date the rating are missing because the data was missing. Just like the correlation that rating didn't correlate with any variables.

5. Are the rating affect with the number of favorite and retweet count?



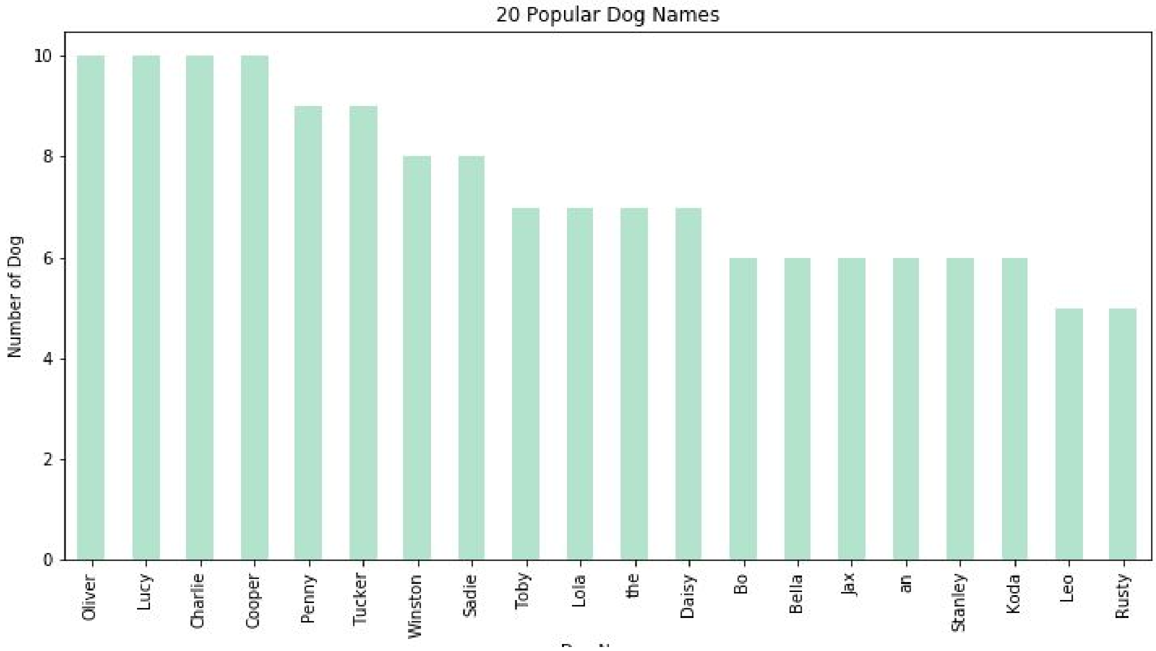
From all data, we found that the higher the rating the higher the count (favorite and retweet). From that plot we also know that count in favorite alway higher than retweet.

6. How much each algorithm predict the picture is dog?



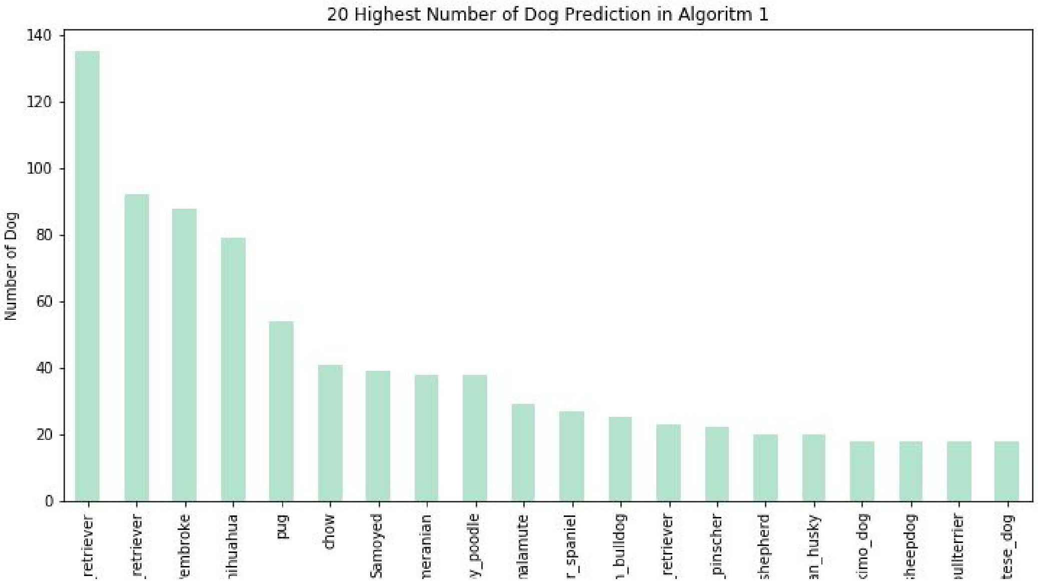
P2 predict picture dog large than p1 and p3. The smallest predicted is dog come from p3.

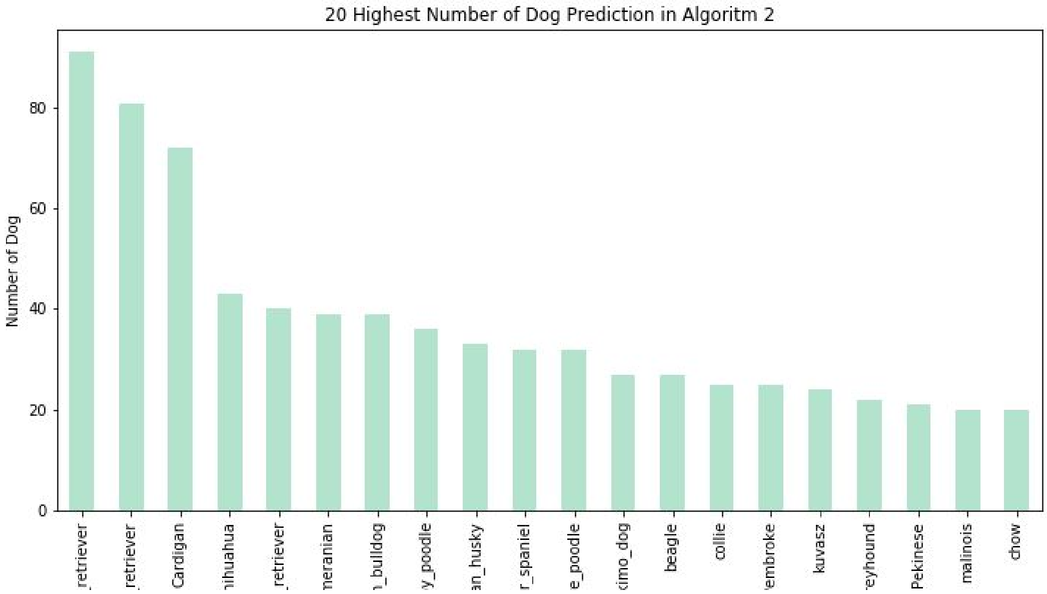
7. What are the most popular dog names?

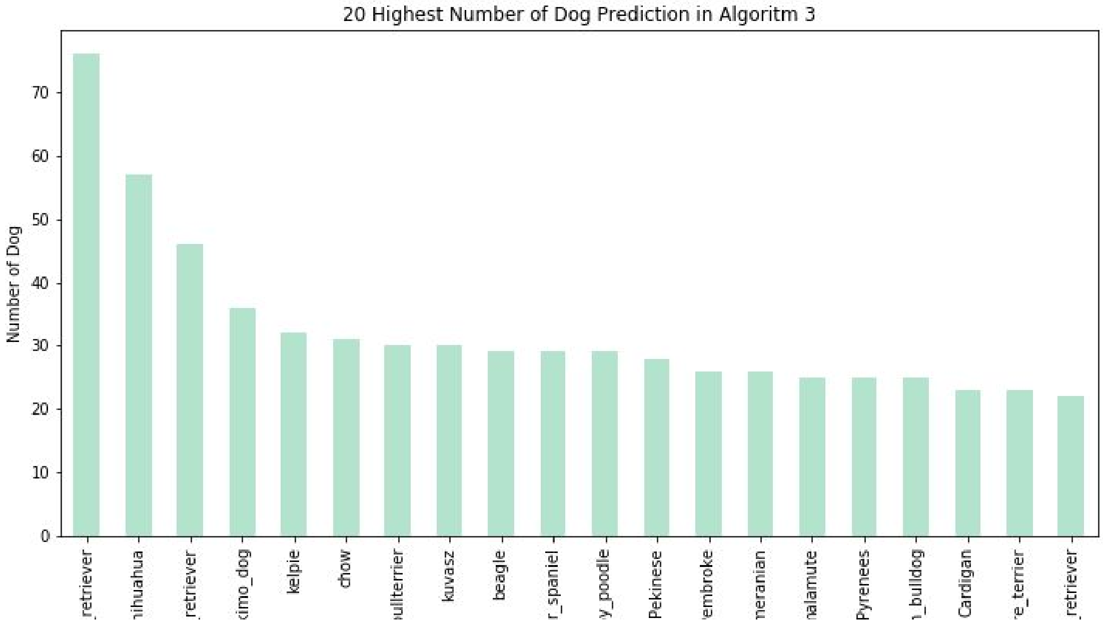


"Oliver", "Lucy", "Charlie", and "Cooper" is the commond dog names in that table.

8. What are the most popular dog predict?







In Algorithm 1, golden retriever are the most popular dog, but in Algorithm 2 and 3, labrador retriever is the most popular dog.

9. What are the most popular dog predict when all algorithms predict the same dog?

When I select data with same answer in at least 2 Algorithm, there are just 1 data, that is predict mailot and exist in algorithm 1 and 2.