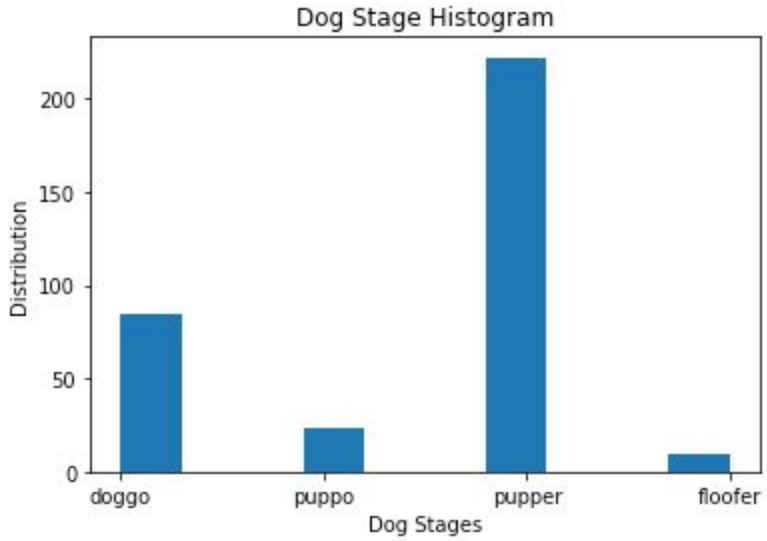
The insights described below are based on the dataset that is included in the file `twitter_archive_master.csv`.

Insights

1. What I first noticed when I started analyzing the dataset was that there are quite a high number of records without a dog stage. I still wanted to see the distribution of dog stages on the records and as can be seen below there are clearly more puppers than any of the other dog stages.



2. The second thing I like to do is to run the .describe() function because it gives me a good idea about the statistics of the numerical values. For example, looking at the numerator and denominator statistics I can see that most there are still some outliers but for the vast majority the values are within the expected ranges. As another example, I can see that from the 3 provided predictions (p1, p2, p3), p1 is the one that has values that will be more helpful because the ranges confidence in the result are considerably higher than the range of confidence that p2 and p3 are providing.

	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	img_num	p1_conf	p2_conf	p3_conf
count	2.115000e+03	2115.000000	2115.000000	2115.000000	2115.000000	2115.000000	1992.000000	1.992000e+03	1.992000e+03
mean	7.363162e+17	12.245863	10.501182	2498.179196	8276.905910	1.133333	0.593960	1.344581e-01	6.021890e-02
std	6.706619e+16	40.290984	7.103397	4390.793427	12079.214074	0.613009	0.271928	1.007064e-01	5.086737e-02
min	6.660209e+17	0.000000	10.000000	0.000000	0.000000	0.000000	0.044333	1.011300e-08	1.740170e-10
25%	6.766157e+17	10.000000	10.000000	554.000000	1828.000000	1.000000	0.362903	5.401683e-02	1.616933e-02
50%	7.094095e+17	11.000000	10.000000	1212.000000	3778.000000	1.000000	0.587635	1.174550e-01	4.950530e-02
75%	7.871428e+17	12.000000	10.000000	2850.500000	10300.000000	1.000000	0.845599	1.952647e-01	9.157912e-02
max	8.924206e+17	1776.000000	170.000000	77917.000000	156349.000000	4.000000	1.000000	4.880140e-01	2.734190e-01

3. Another thing that I was curious to see visualized was the histogram of the dog rates. I first did the histogram of all the values. The second histogram includes only the tweets that have the highest number of retweets. And finally the third histogram includes only the tweets that have the highest numbers of favorites. Even though all three have kind of the same distribution, I can see that the distribution of the histograms with the favorites and retweets have less rates that are less than 10.
   Figure 1.

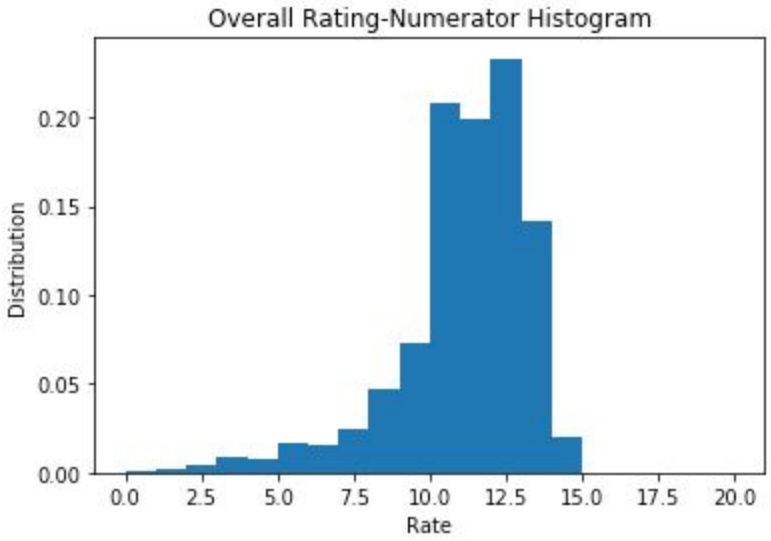


   Figure 2.

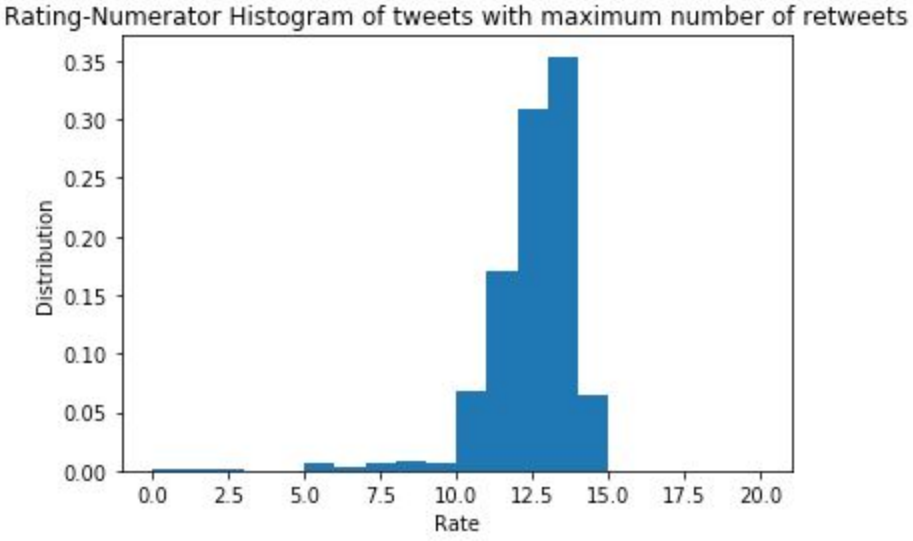
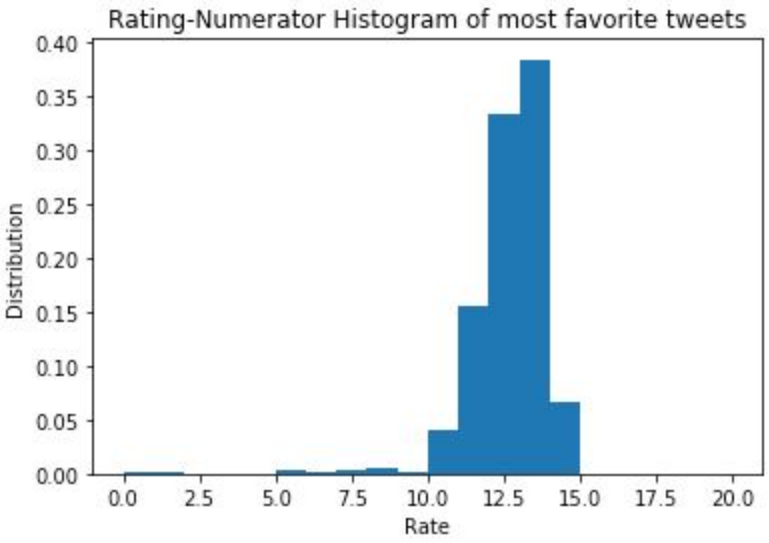


   Figure 3.


Rating-Numerator Histogram of most favorite tweets
Distribution
Rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0