Insights Report: WeRateDogs Twitter Archive

Overview

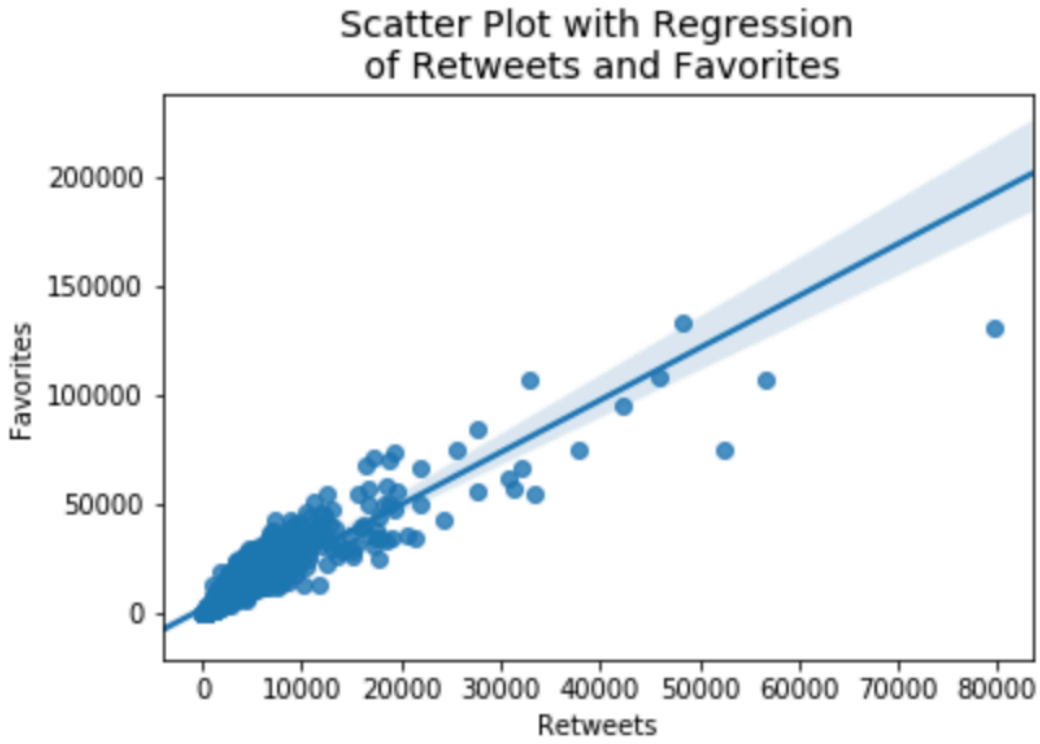
Analysis of the tweet archive of Twitter user @dog_rates (WeRateDogs) produced some interesting insights, some of it predictable while others were not so obvious. The major effort of this task was analyzing and cleaning of the data which took over 40 hours of effort. The data provided was fairly dirty, but once cleaned, there were a number of insights that were found.

Insight 1 – Tweets that are 'favorited' have a high correlation to being 'retweeted'

The first thing checked was to see if any of the numerical data was correlated.

	rating	retweet_count	favorite_count	img_num	p1_conf	p1_dog
rating	1	0.018	0.016	-0.00016	-0.0027	-0.047
retweet_count	0.018	1	0.91	0.11	0.026	0.025
favorite_count	0.016	0.91	1	0.14	0.057	0.065
img_num	-0.00016	0.11	0.14	1	0.14	0.052
p1_conf	-0.0027	0.026	0.057	0.14	1	0.023
p1_dog	-0.047	0.025	0.065	0.052	0.023	1

- Retweets and Favorites is highly correlated with a correlation coefficient of 0.91.
- This was even more obvious when shown with a regression graph.

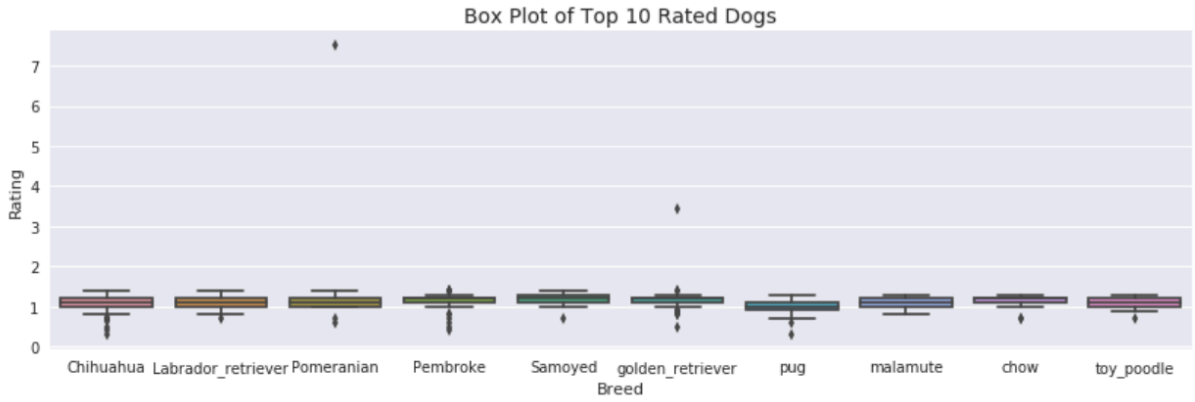


<u>Insight 2 – Samoyed, Golden Retriever, & Pembroke are the three highest rated dogs by mean (when removing of outlier data)</u>

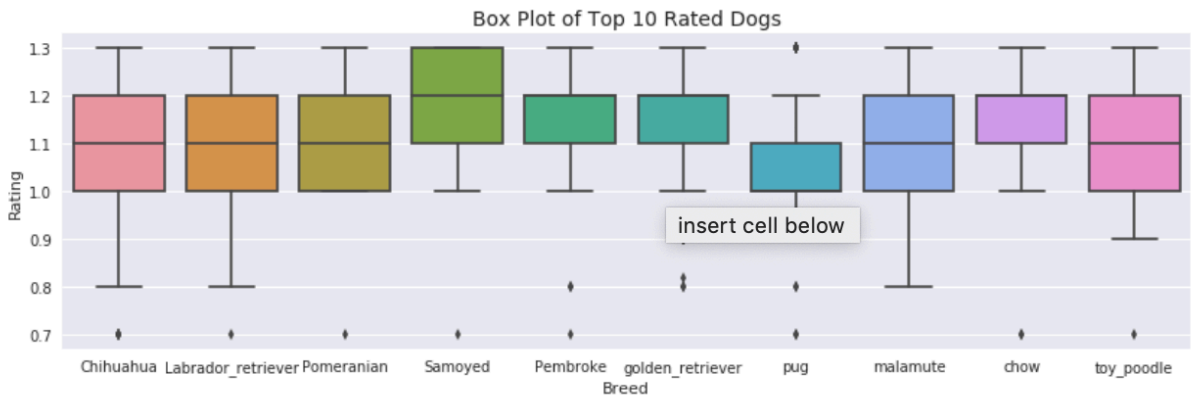
- In order to ensure sufficient data for insights, only the top 10 rated dogs (by number of ratings) was used.
    - Most rated (by count): Golden Retriever with 156 ratings.
    - Lease rated (by count): Malamute with 33 ratings.

	rating			
	count	mean	min	max
p1				
Pomeranian	41	1.253659	0.6	7.500000
Samoyed	42	1.169048	0.7	1.400000
golden_retriever	156	1.167607	0.5	3.428571
Pembroke	94	1.142553	0.4	1.400000
chow	48	1.141667	0.7	1.300000
Labrador_retriever	106	1.119811	0.7	1.400000
toy_poodle	50	1.100000	0.7	1.300000
malamute	33	1.087879	0.8	1.300000
Chihuahua	90	1.049293	0.3	1.400000
pug	62	1.024194	0.3	1.300000

- Pomeranian is the highest rated dog, but there appears to be some extreme outlier data with one of the ratings being 7.5.
- We can look at this visually using a box plot.



- Checking for data, the highest rating within 2 standard deviations is 1.3. Let's remove all ratings outside of a 2 standard deviation and see how the box plots change.



- Now let's see how this changed the rankings:

	rating			
	count	mean	min	max
p1				
Samoyed	41	1.163415	0.7	1.3
golden_retriever	152	1.154067	0.8	1.3
Pembroke	87	1.152874	0.7	1.3
chow	48	1.141667	0.7	1.3
Labrador_retriever	105	1.117143	0.7	1.3
Pomeranian	38	1.102632	0.7	1.3
toy_poodle	50	1.100000	0.7	1.3
malamute	33	1.087879	0.8	1.3
Chihuahua	85	1.072941	0.7	1.3
pug	60	1.043333	0.7	1.3

- With the outliers removed, Pomeranian moves from #1 to #6 in the rankings.
- Samoyed, Golden Retriever, & Pembroke are the three highest rated dogs by mean.

Insight 3 –Without being provided follower count, we can still extrapolate that follower count has significantly increased between 2015-2017 based on the mean increase of 'favorites' and 'retweets' per tweet.

year	rating	favorite_count	retweet_count
2015	0.964505	2185.281081	897.841441
2016	1.076850	6799.279781	2221.434973
2017	1.230769	19889.819398	4470.397993

- Mean Favorite and Retweet counts double every year which can be interpreted as follower count going up significantly year-over-year.