Data Wrangling - Act Report

by Katrin Haller/ September 2018

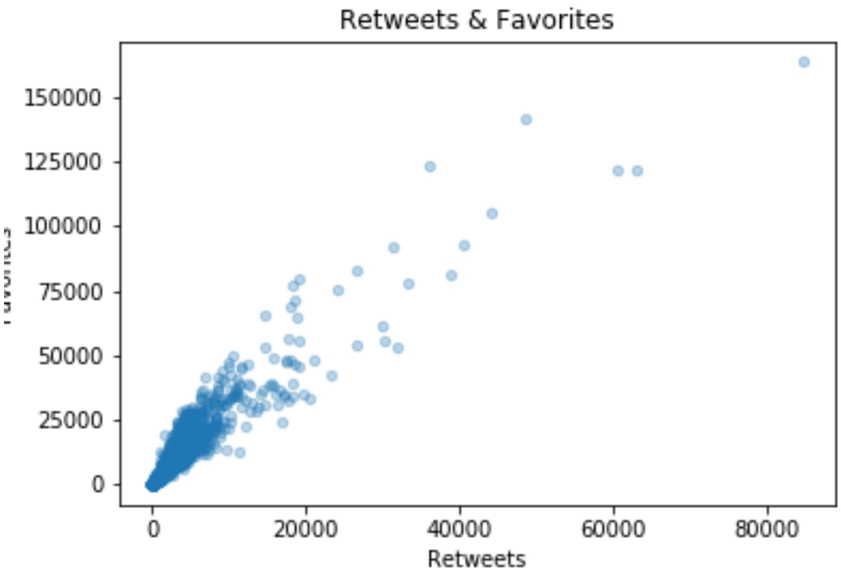
I analyzed Data from the Twitter Account WeRateDogs after I gathered, cleaned and stored the data. While analyzing the data I first tried to figure out interesting insights.

I started visually by looking at my dataframe and made up my mind for interesting questions. Therefore in the main statistics I found outliers for the ratings, but also that the most ratings lie between 10 and 12 with a mean rating of 12. The confidence level for the neural network prediction of the pictures was 0.8, which is really good. I also saw that on every third favorite follows a retweet.

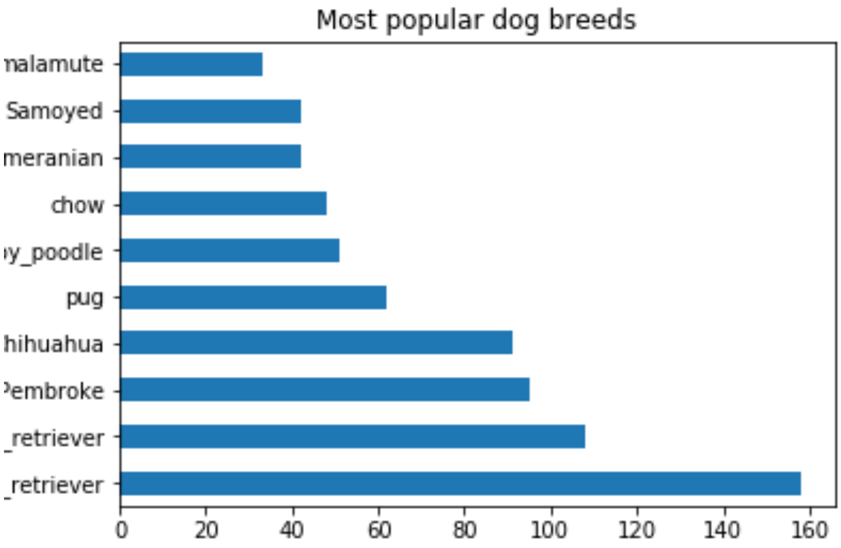
---	tweet_id	rating_numerator	rating_denominator	img_num	conf_level	retweets	fa
count	1.994000e+03	1994.000000	1994.000000	1994.000	1994.000	1992.000000	1
mean	7.358508e+17	12.280843	10.532096	1.203109	0.464991	2693.973896	8
std	6.747816e+16	41.497718	7.320710	0.560777	0.339470	4773.777838	1
min	6.660209e+17	0.000000	2.000000	1.000000	0.000000	12.000000	8
25%	6.758475e+17	10.000000	10.000000	1.000000	0.140466	598.000000	1
50%	7.084748e+17	11.000000	10.000000	1.000000	0.459130	1299.500000	4
75%	7.877873e+17	12.000000	10.000000	1.000000	0.776387	3088.250000	1
max	8.924206e+17	1776.000000	170.000000	4.000000	0.999956	84467.000000	1

I decided to have four sections for getting interesting insights. (1) Retweets & Favorites (2) Dog Breed (3) Dog Names
(4) Ratings

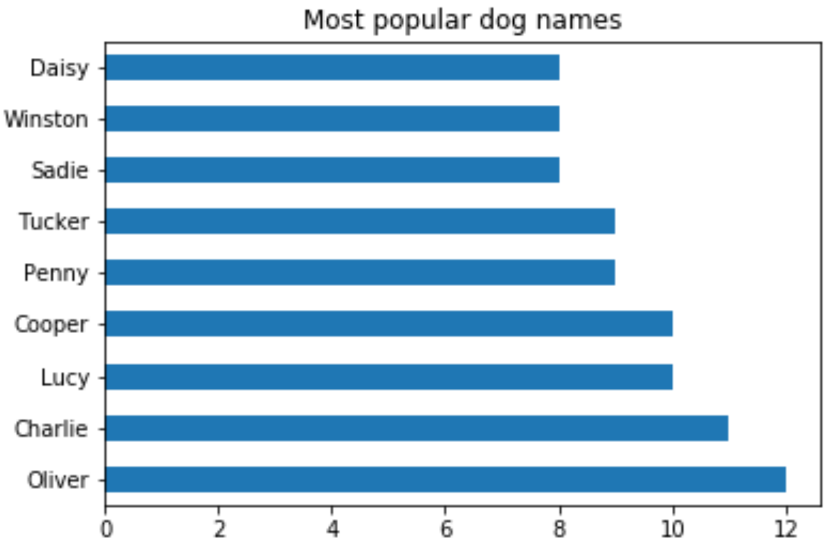
For all of these steps I created visuals as possible to have a better understanding. As expected the retweets and favorites have a strong correlation as you can see here.



My master dataframe also shows 113 different dog breeds with Golden Retriever far at the top.



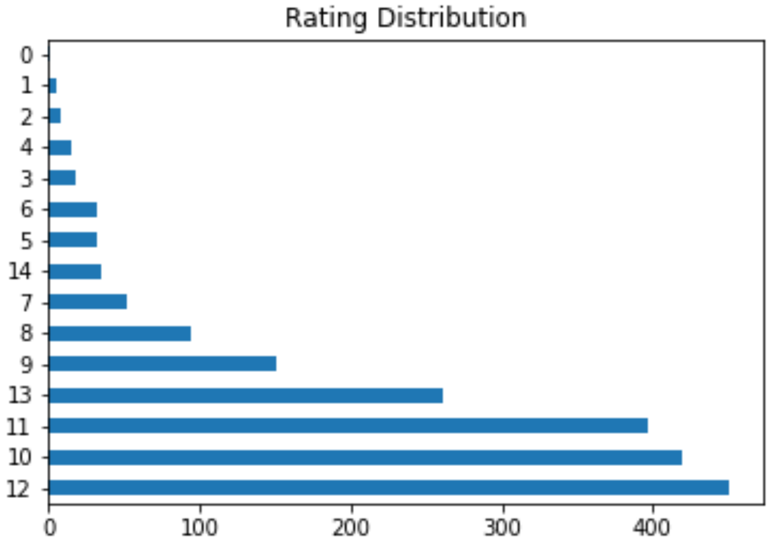
We also have 922 different dog names, which seems pretty much when you consider to have about 1900 dogs in the dataframe. Then I found that 582 dogs are listed without names (None) and a lot of names just given ones. The most popular dognames here were 1.Oliver, 2. Charlie and 3.Lucy and Cooper, 4.Penny and Tucker, 5.Daisy, Winston and Sadie.



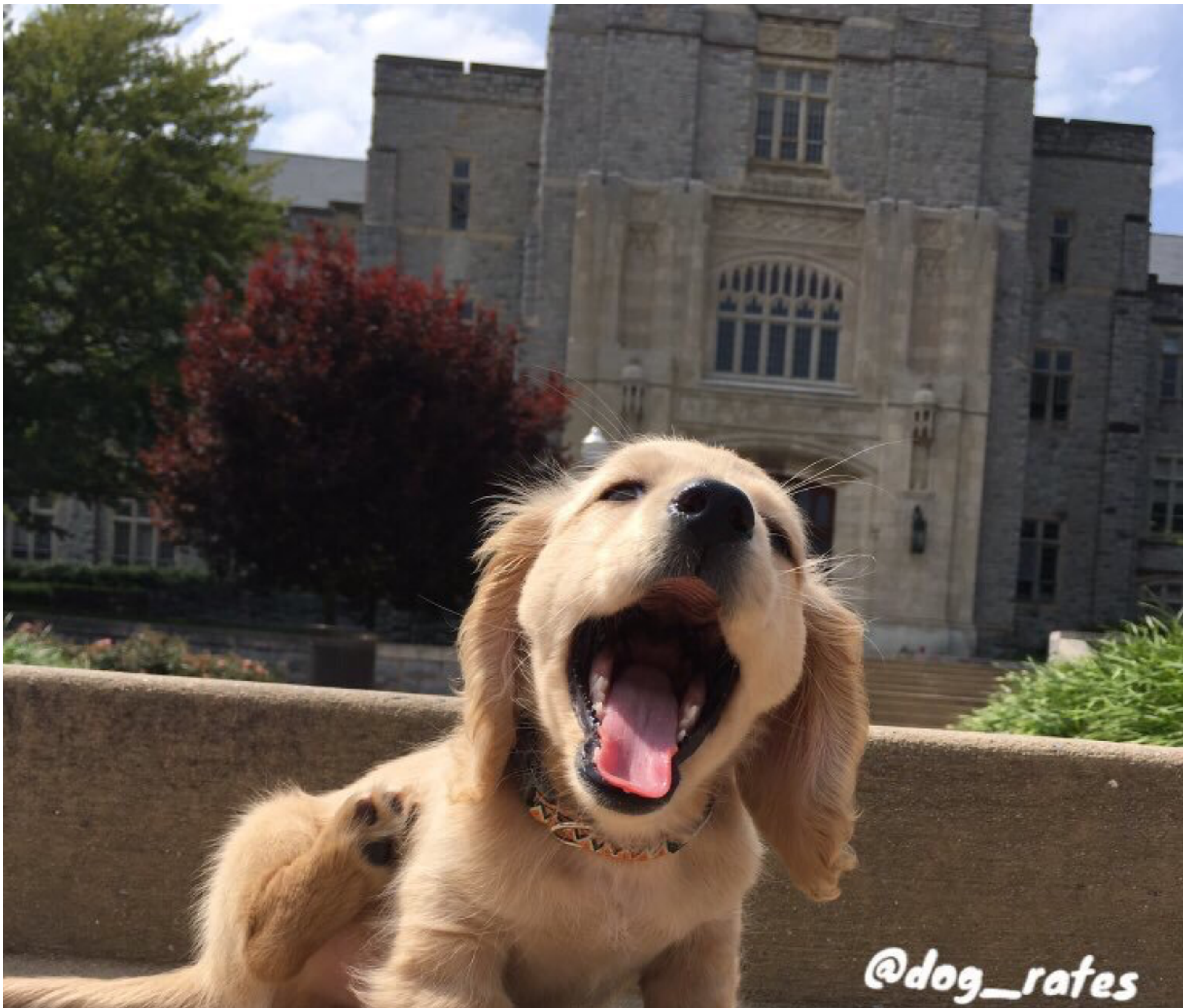
In the Ratings section I found that the mean rating for the dog stages is 12 for puppo, but closely followed by doggo and floofer with 11.8 and finally pupper with 10.7.

dog_stages	**value**
doggo	11.888889
floofer	11.875000
pupper	10.726415
puppo	12.043478

Nearly the same results like for the ranking of dog stages I got for the dog stages and their retweets. Here was doggo on top with 7163 retweets, followed by puppo with 6913, floofer with 4575 and pupper with 2356 retweets. In this plot here you can see the distribution of the most common ratings from WeRateDogs with the rating numerator of 12 on top.



Finally here is a picture from a **Golden Retriever** named **Oliver** - number 1 of dog breed and names.


@dog_rates