INSIGHTS FROM WERATEDOGS

By Bilal Arshad Chaudhry

Dogs, since time immemorial, have been our best and most faithful companions. It is estimated that dogs, being the genetic descendants of wolves, were domesticated around 27,000 to 40,000 years ago through a process of natural selection and human intervention when they started getting attracted to camps and settlements. In time, this relationship became one of the cornerstones of the human civilization and has undergone numerous transformations during its evolution. Dogs were once and still are, in many parts of the world, central to the security of our livestock, food stores and personal property.[1]

As our attachment with these wonderful animals grew, so did our love for sharing stories about them. From the spoken word to papyrus scrolls, we as a species always had a tale to tell when it came to our furry canine friends. Fast forward to the 21st century and we're still talking or sharing content about the dogs in our lives, albeit on a much different medium and at a scale previously unimaginable. This is where the Twitter account, WeRateDogs, comes into the picture. This account is a repository of tweets which contain pictures of people's dogs, along with a rating (from 0 to 10) and a humorous comment. The account has over 6 million followers and has received considerable coverage from various media outlets. And where there is data, there are data analysts, itching to gain some sort of insight from the ocean of information at their fingertips. So, armed with Python's pandas and visual libraries, we will dive into our analysis.

Prepping the Data

In order to conduct this analysis, we'll be gathering data from three sources i.e. an archive of tweets (till August 2017) made available by Udacity, the Twitter API containing additional information regarding individual tweets and, finally, results from an image prediction neural network which classifies the different breeds of dogs based on images in tweets.

A Look at the Statistics

Before all else, one must analyse the descriptive statistics of any cleaned dataset to develop a basic understanding of the variables involved. Of 1910 observations, we can observe that people, in general, tend to disregard the rating system by…

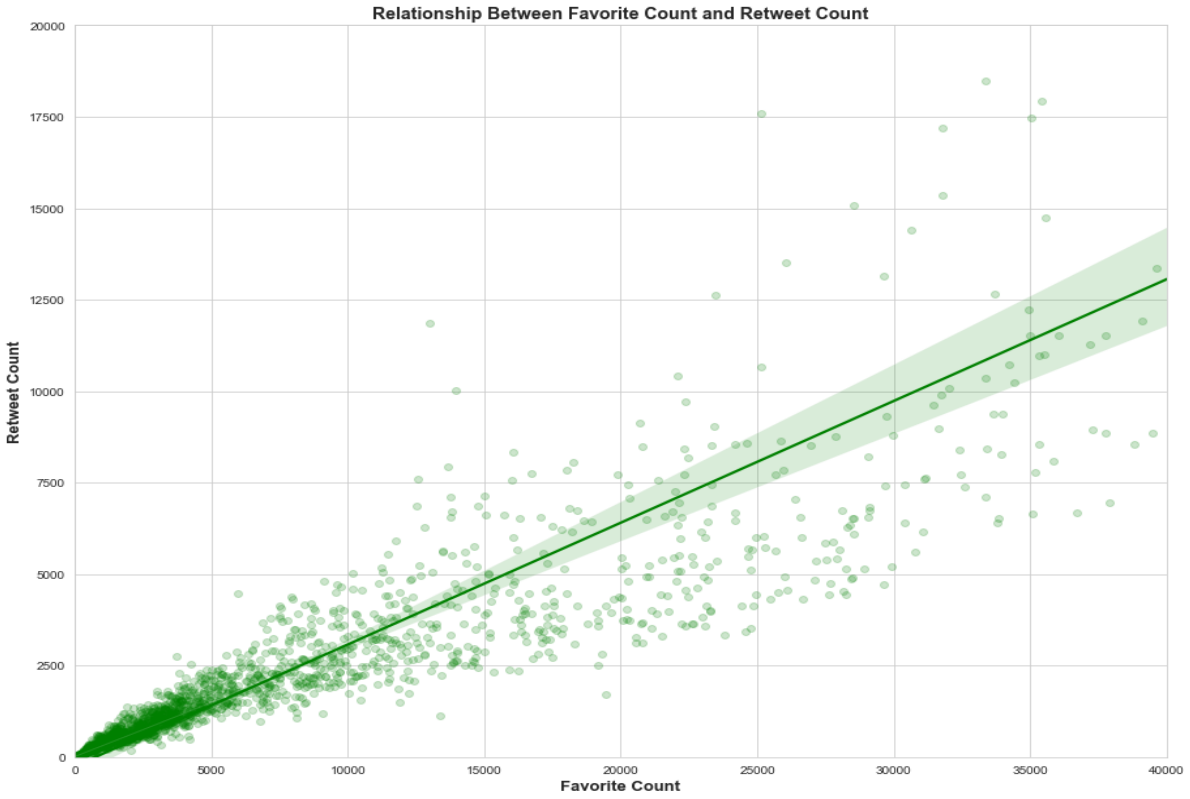
	tweet_id	rating	confidence_lvl	retweet_count	favorite_count	year	month	day
count	1.910000e+03	1910.000	1910.000	1910.000	1910.000	1910.000	1910.000	1910.000
mean	7.351744e+17	11.645	0.463	2514.864	8316.849	2015.841	7.164	16.068
std	6.789534e+16	41.523	0.341	4192.012	11403.997	0.704	4.126	8.947
min	6.660209e+17	0.000	0.000	16.000	81.000	2015.000	1.000	1.000
25%	6.755190e+17	10.000	0.133	611.000	1861.500	2015.000	3.000	8.000
50%	7.073066e+17	11.000	0.457	1298.000	3904.500	2016.000	7.000	16.000
75%	7.866477e+17	12.000	0.776	2918.000	10309.250	2016.000	11.000	24.000
max	8.924206e+17	1776.000	1.000	79515.000	132810.000	2017.000	12.000	31.000

…over-estimating their dog's rating. The mean rating is 11.645, which means that a large proportion of the users rated their dog 10 or above with 1776 as the highest value. The mean value for retweets is approximately 2515 while the mean for favorited tweets is approximately 8317. What is interesting to note is that the mean value for prediction confidence levels is 0.46 or 46%; indicating a need to improve the prediction CNN.

[1] The History of Dogs as Pets: https://abcnews.go.com/Lifestyle/history-dogs-pets/story?id=41671149

Retweets and Favorites

Moving on from the descriptive statistics, we can turn our attention to the relationship between retweets and favorites.



From the plot above, we can clearly observe a strong positive correlation between favorite counts and retweet counts. With a Pearson Coefficient / **r** value of **0.906**, it can safely be assumed that there is a high probability of tweets with high retweet counts also having high favorite counts. While the majority of tweets have below 5000 retweets and 2500 favorites, there also exists a significant number of tweets with counts way above these values. This shows the high level of involvement exhibited by the followers of WeRateDogs.

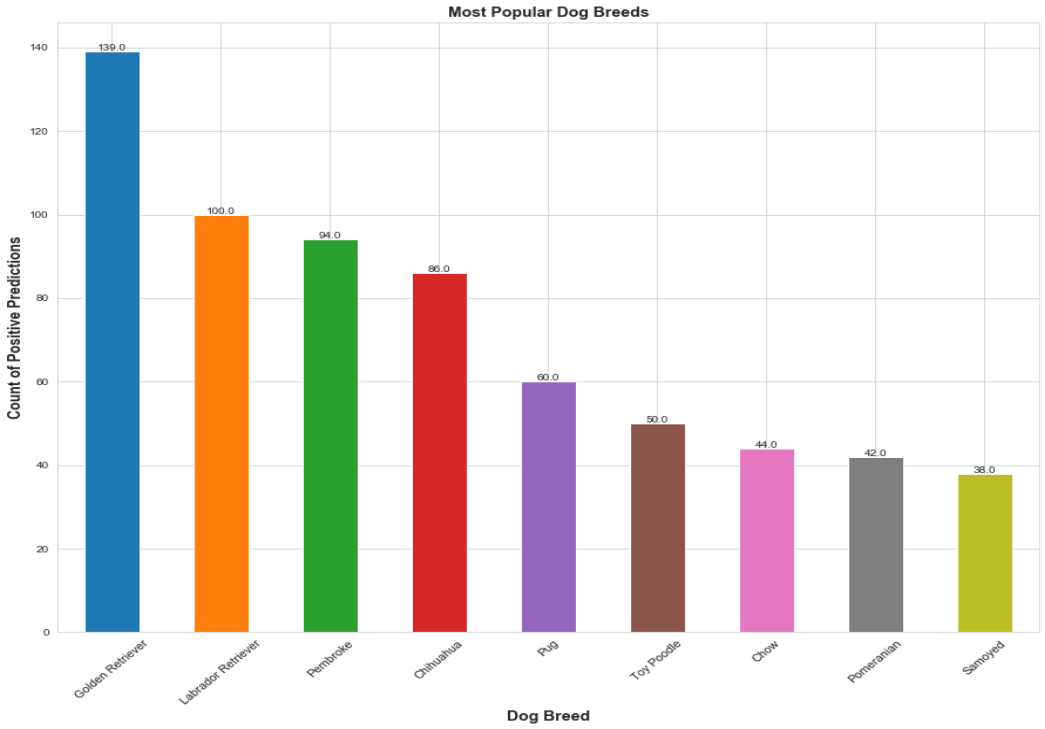
The Most Popular Breeds

A storehouse of data as large as WeRateDogs, provides us with the perfect opportunity to assess which breeds tend to be the most popular. If we take a look at the sum of retweets and favourite counts for the most popular breeds, it becomes apparent that the Golden Retriever and the Labrador Retriever take the top two spots with 460,827 and 341,361 total retweets. Additionally, these two breeds have 1,602,374 and 1,092,075 likes respectively. While the breeds with the lowest scores for retweet and favorite counts are the Japanese Spaniel, Groenendael and Entlebucher with 471, 553 and 706 retweets respectively.

breed	retweet_count	favorite_count
Not Dog	691478	2002314
Golden Retriever	460827	1602374
Labrador Retriever	341361	1092075
Pembroke	273271	969613
Samoyed	158426	472290
Chihuahua	153241	543649
French Bulldog	143712	515823
Chow	116202	406003
Toy Poodle	112798	331377
Pomeranian	112171	312957

breed	retweet_count	favorite_count
Entlebucher	706	2678
Groenendael	553	2313
Japanese Spaniel	471	1362

The plot below shows that the most popular breed, by a wide margin, is the Golden Retriever with 139 positive predictions. Second and third place go to the Labrador Retriever and the Pembrooke with 100 and 94 predicions respectively. Chihuahua's, Pugs and Toy Poodles make up fourth, fifth and sixth places with 86, 60 and 50 predictions. Finally, Chows, Pomeranians and Samoyeds take seventh, eighth and ninth place with 44, 42 and 38 predictions.



However, these results come with a caveat; that being the results of these predictions are only as good as the CNN's algorithm. So, as before, these results reflect the outcomes of the current CNN and may change with a code revision for improved outcomes.

Thus, we've managed to scratch the surface of this topic with the data available to us. Over time, this data can be revisited to glean greater and deeper insights or, perhaps someday, yours truly will attempt to program a CNN of his own to improve the image prediction outcomes. All in all, this exercise was an eye opener as far as the limitless possibilities of data analysis are concerned. I can't wait to get started on the next one!