WeRateDogs Twitter Archive – Data Analysis

I started to analyze the wrangled WeRateDogs Twitter data and associated image predictions. The initial insights here may provide ideas for more in-depth analyses.

Attempt at modeling favorite count based on stage of a dog's life

The first question I wanted to explore was whether we could model the favorite count based on the stage of a dog's life. As defined in The Dogtionary, the stages of a dog's life include pupper, poppo, and doggo. This type of model could be useful if I wanted to post statuses about my dog as it grows up, and understand how popular it may be over time.

The linear regression was done using the statsmodels module's OLS method. Unfortunately, the R-squared value was very low at 0.039, so a linear model is not a good fit for this data. The p-values for puppo and pupper dog stages were close to zero though. This suggests that these dog stages have some effect on favorite count.

Some dogs did not have a dog stage, and I labeled these as "other". It may be interesting to find the missing dog stage data, then try fitting the model again.

Popularity of Floofer vs Non-Floofer

Next, I looked at the popularity metrics (rating, favorite count, and retweet count) of dogs based on their "floofer" status. A dog can either be a floofer or a non-floofer, depending on how fuzzy and fluffy they are. See the data table below for reference.

Floofer Popularity

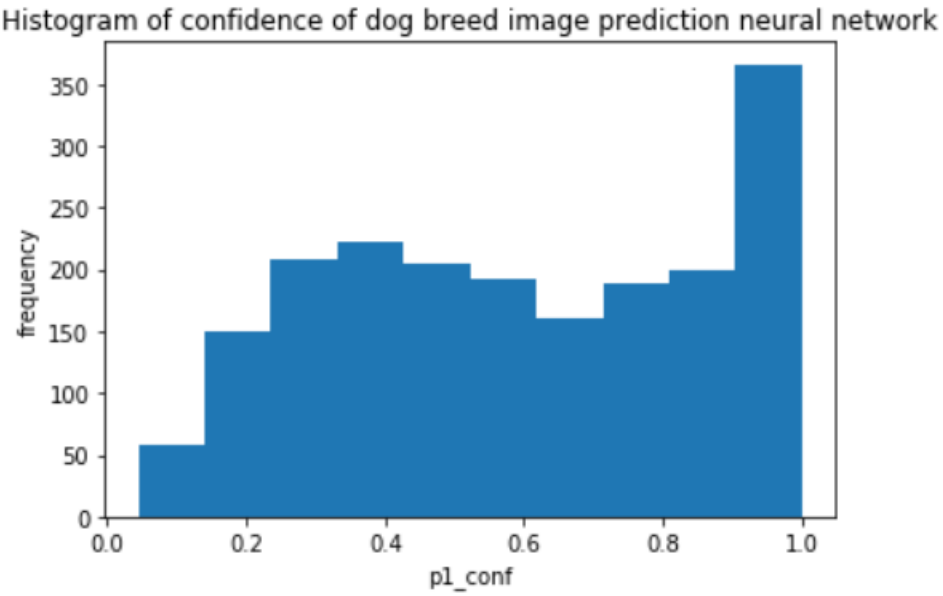
	Rating (mean)	**Rating (std dev)**	**Favorite Count (mean)**	**Favorite Count (std dev)**	**Retweent Count (mean)**	**Retweet Count (std dev)**
Non-Floofer	10.6	2.1	8,117	11,886	2,431	4,286
Floofer	11.8	1.0	10,434	9,168	3,480	4,445

Floofers had similar rating as non-floofers. It is less clear whether floofers had different favorite count or retweet count as non-floofers, as the standard deviation for these metrics is the same order of magnitude as the mean.

In a future analysis, I could investigate the floofer data and get a better understanding of the large spread in favorite count and retweet count.

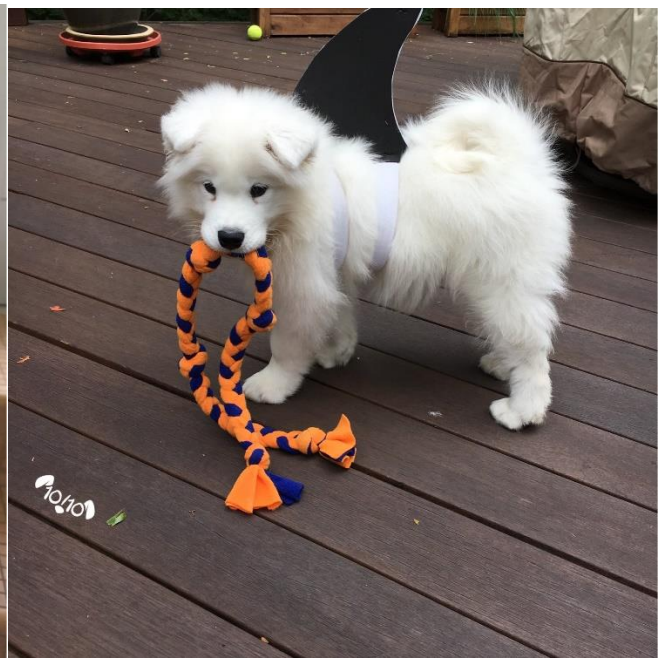
Prediction of dog breed based on image using neural network

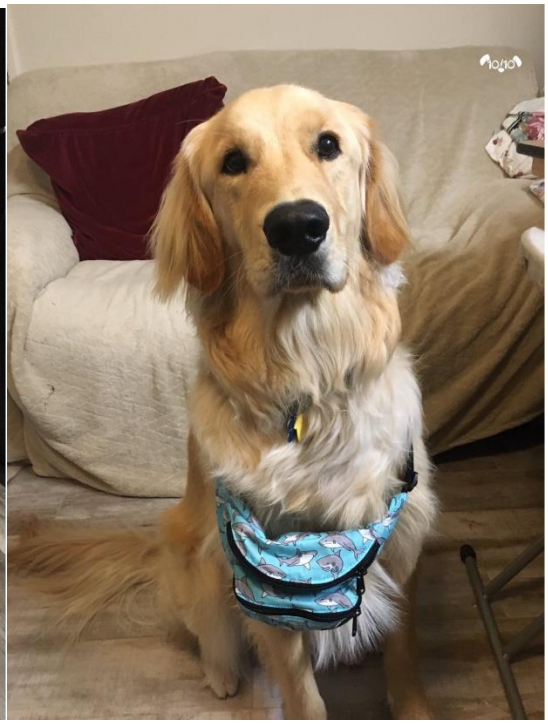
Finally, I looked at the image prediction data for dog breed. See the histogram below. Here, “p1_conf” is the confidence level of the prediction.



The histogram shows that the dog breed image prediction neural network had varying degrees of confidence. The bin with highest confidence also had highest frequency, which is great to see.

Dog breeds where the image prediction most frequently had high confidence (over 95%) were pug, samoyed, pembroke, and golden retriever. It would be interesting to understand whether these breeds have common features that make it more likely for the model to identify them. Below are images of these breeds that the model correctly identified.





Wrap-Up

The notes above are a starting point. Future analyses could do more data exploration and more accurate modeling. There may also be a data source that could fill in the missing dog stages.

Thank you for reading!