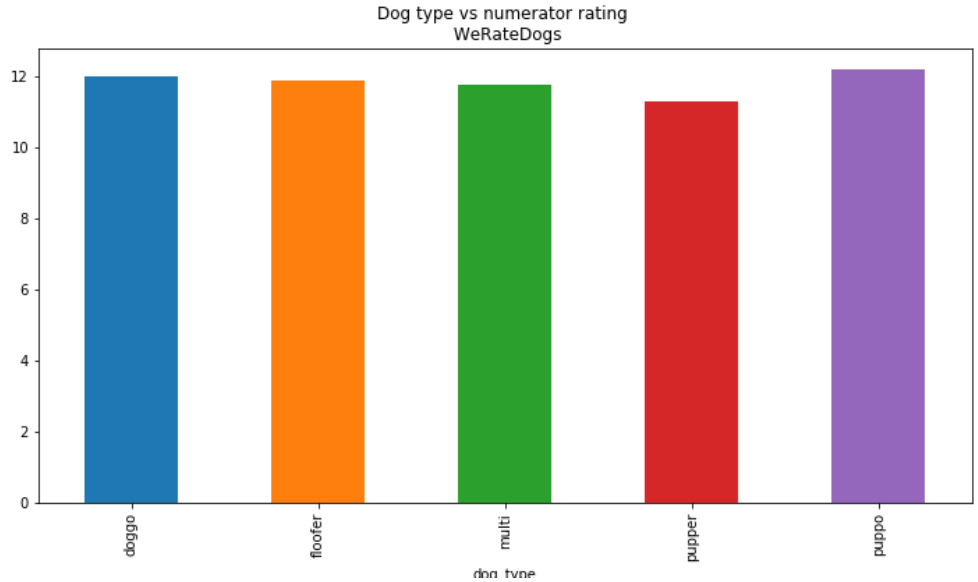
Act report

For this section I've decided to focus on three separate insights:

What dog type has the highest average rating?

- I've used t_archive_clean_long df. I've grouped dog_types and aggregated the mean rating_numerator. puppo dog tag seem to be the most liked with an average of 12.8. I've used a barh plot to display the results.



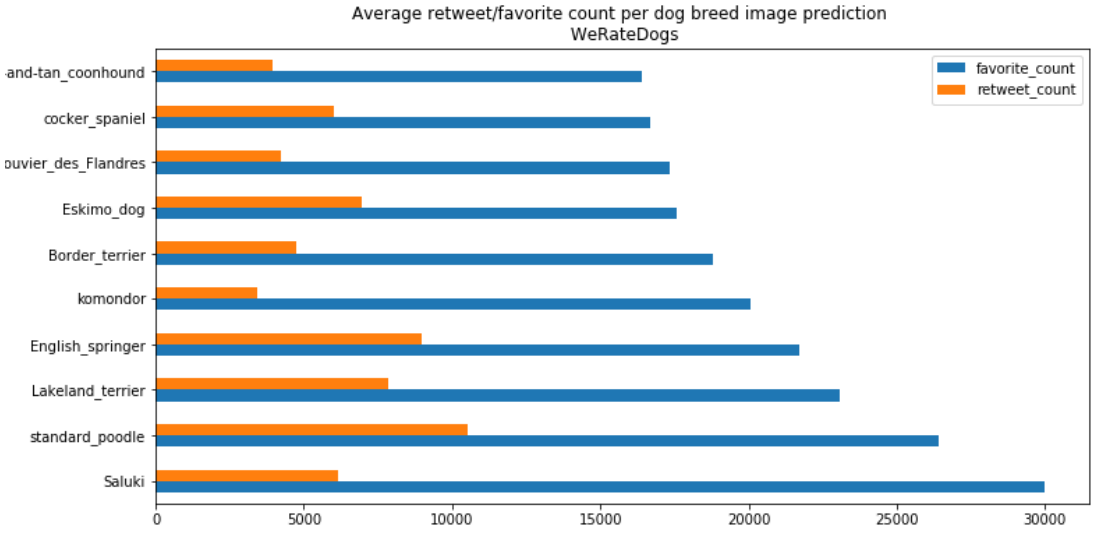
What are the top 10 dog names?

- Also used t_archive_clean_long df. I've used a simple value_counts to figure out if there was a popular dog name. Dog names appear to be part of a long tailed distribution. The most frequent name 'Lucy' only represented 0.79% of all values.

What dog breed has the highest tweet/retweet count average?

- Here twitter_archive_master_long df was used. I've created the max_value function. In order to answer this question we required a filter on all tweet_id isolating the max confidence value per tweet(conf variable). Once that's is calculated the equivalent value for other Series (breed, retweet_count and favorite_count) is retrieved using an OrderedDict(). The result dataframe was the highest predicted model value per dog_breed identified in each tweet image. I've grouped vales per dog_breed and aggregated the average favorite_count and retweet_count. The top 10 dog breed by

favorite_count are presented in the plot below.



Based on the master data. Saluki (image below) has the highest average `favorite_count`. In the other hand the standard poodle has the highest average `retweet_count`.
Further investigation on the model predictive ability vs breed would be valuable to answer the above question completely. A selective bias of the model vs a particular breed could be a possible effect and skew the results.