# Udacity Data Wrangle Module
# Project – Wrangling Twitter Data (WeRateDogs)

***Data Analysis***

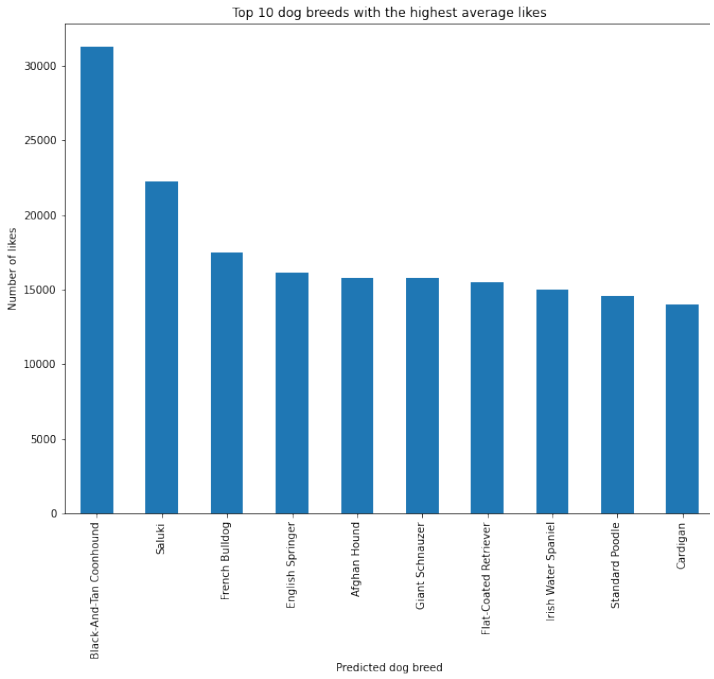
May 22, 2020
Author: Jiayi Zhang

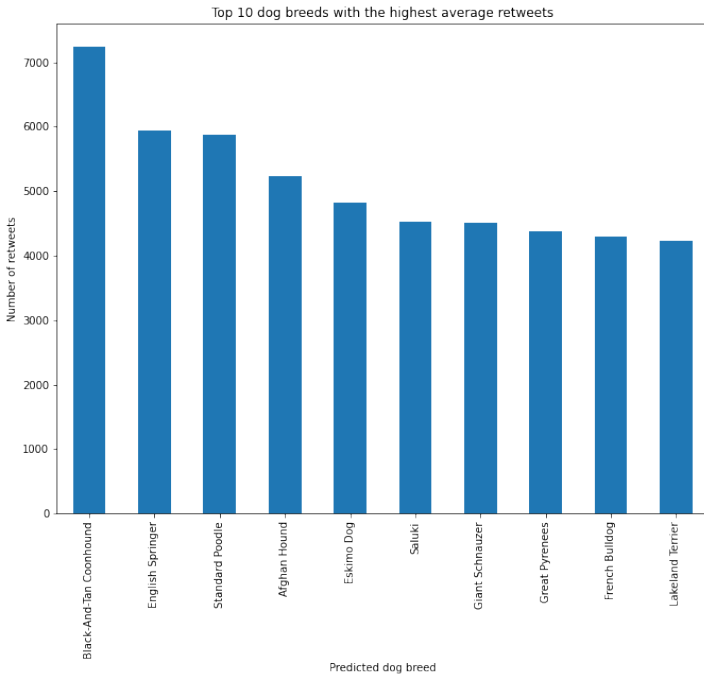
**Motivation**: Data analysis performed on my cleaned dataset (*WRD_clean*) will be described in this section

I have decided to perform my analysis based on these 5 topic questions:

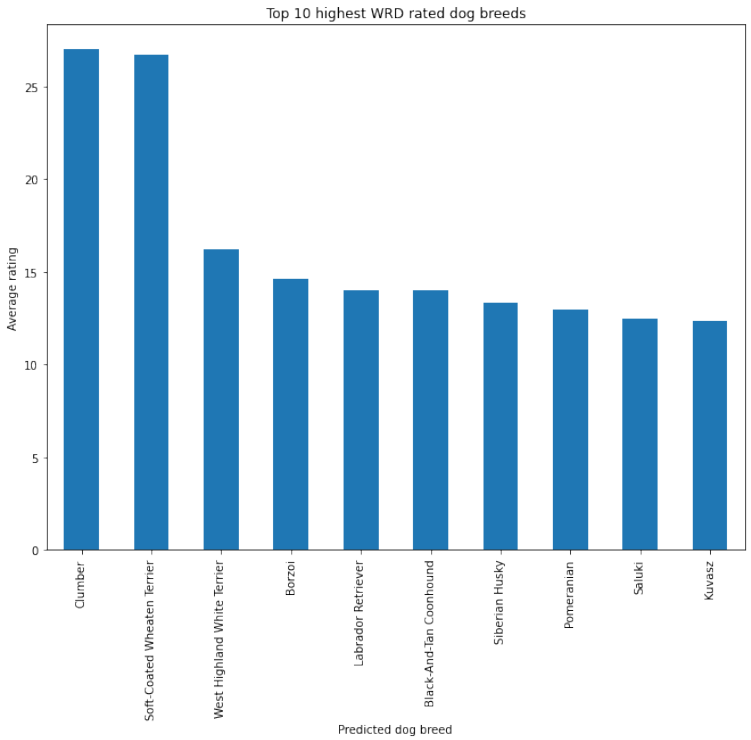
1. Which breed of dog will receive the most retweets and likes?
2. Which breeds are most and least submitted?
3. Does the machine learning algorithm predict certain breeds with higher confidence than others? Which breeds are the easiest and hardest to predict?
4. Does the relative age of the dog (*dog_stage*) affect the confidence of the ML algorithm?
5. What are the trends in retweets and likes over time?

**Analysis 1:**



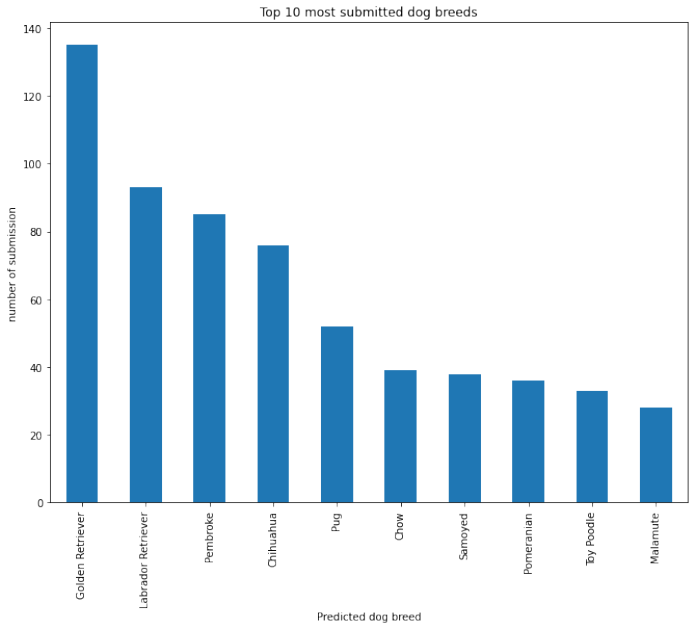



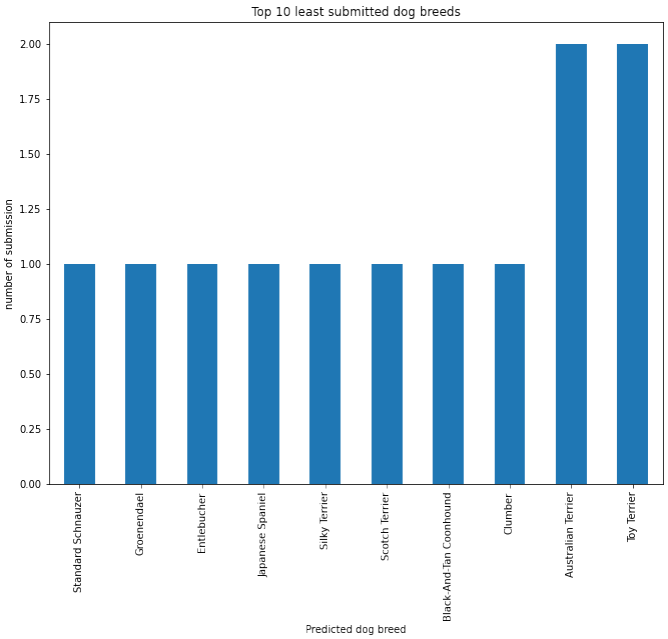

Based on the average number of retweets and likes grouped by dog breed, it appears that the black-and-tan coonhound receives the highest number of likes and retweets (by quite a significant amount as well). Let's compare these results with the WRD ratings (which I find to be a rather subject and arbitrary measure)




Although not the highest rated (by WRD standards), the black-and-tan coonhound is amongst the top 10 highest rated dog breeds.

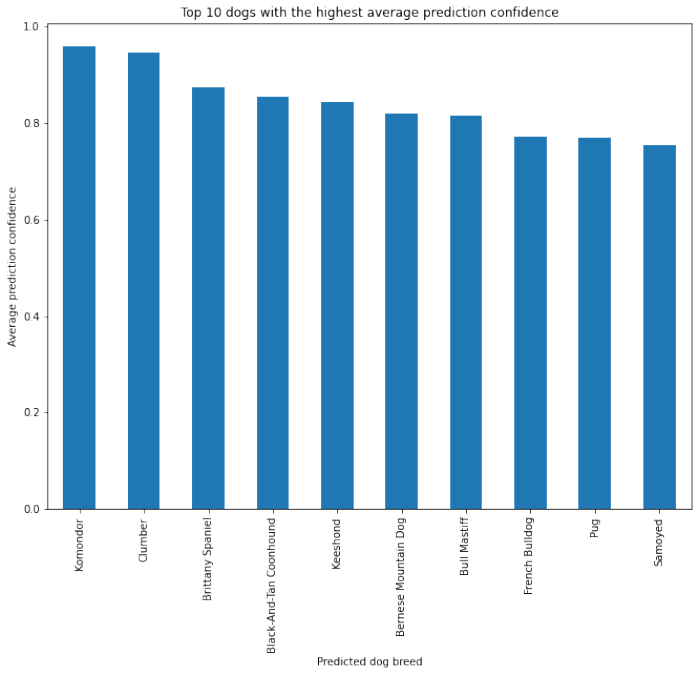
**Analysis 2:**

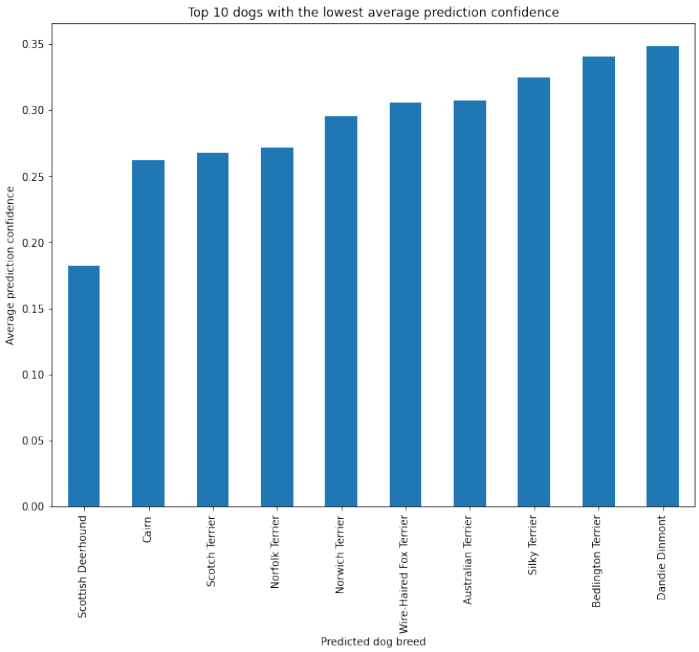



This one is a relative straightforward comparison; which dog breeds are most and least submitted? The idea behind this comparison was to gauge which dog breed was most popularly amongst dog owners. Golden retrievers take home this title with close to 140 submissions, while 7 breeds are least popular with only one submission each.

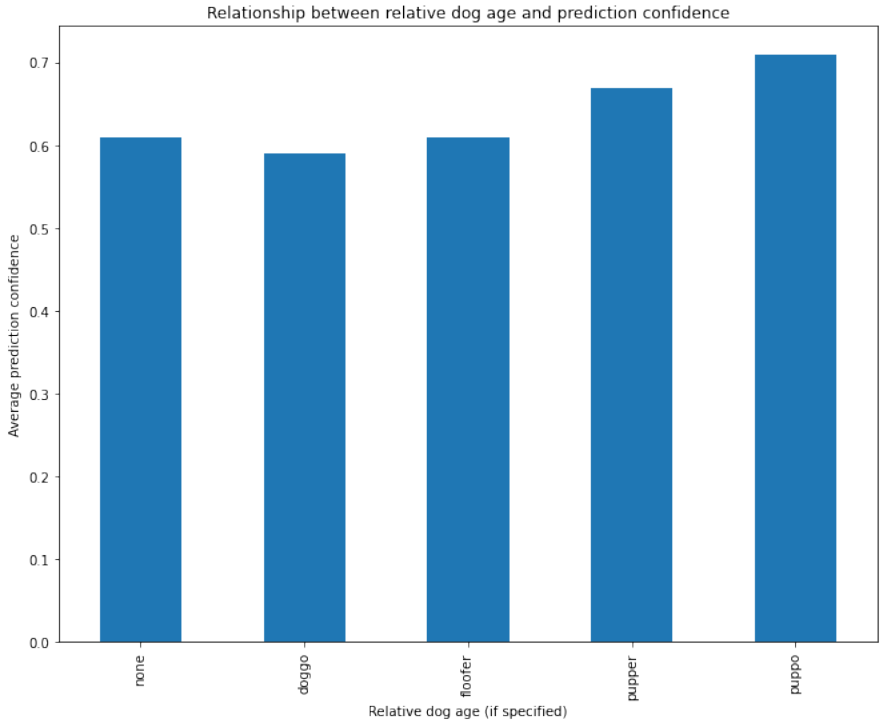
**Analysis 3:**



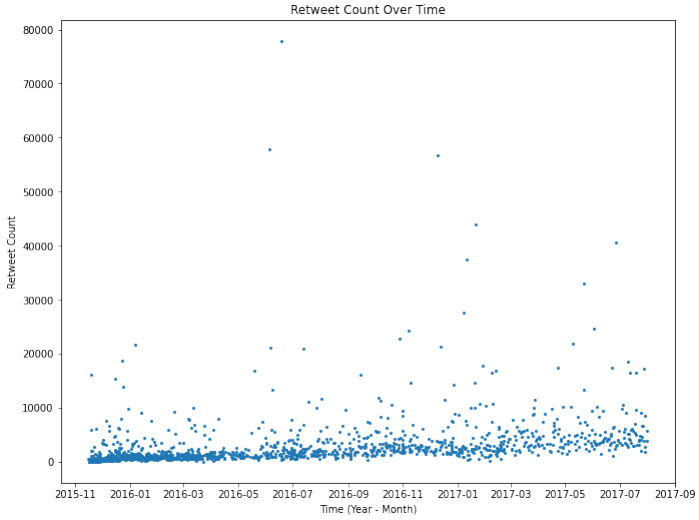

Because the breed of each dog was a prediction from a separate machine learning algorithm, I wanted to see which dog breeds were predicted with the highest and lowest average confidence. Based on these charts, it would appear that Komondors had the highest prediction confidence, coming in at an average of ~0.95 (and if we look at a picture of a Komondor, we can tell that this breed has a very unique and distinguishable look), while the Scottish Deehound was the least confident prediction with an average of ~0.18. It should be noted though the confidence of the prediction depends on the model, so it could be the fact that there was a lot more training data for Komondor than Scottish Deerhound for the algorithm, which would naturally affect its predictions.

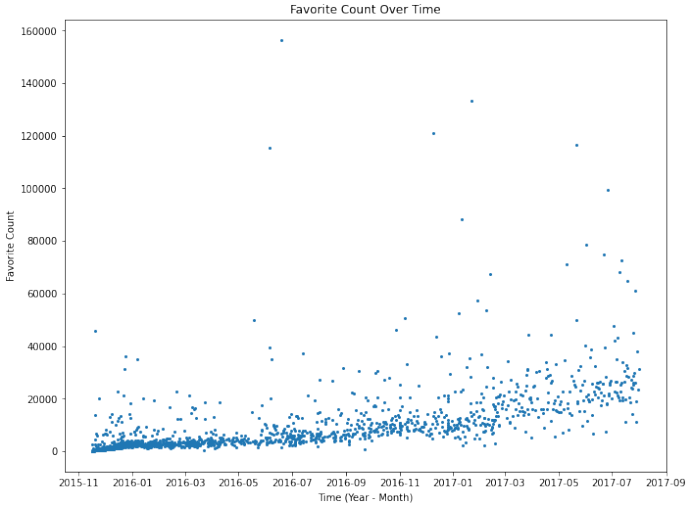
**Analysis 4:**



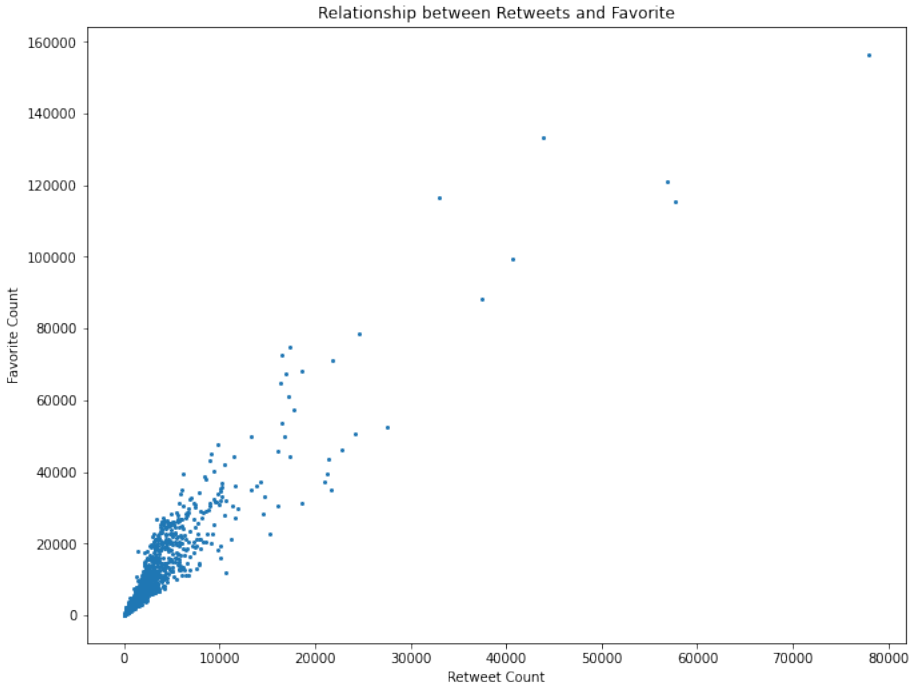
Similar to the previous analysis, I wanted to see whether or not the relative age of the dog affected the confidence of the prediction. Based on our data, there seems to be a positive correlation between age and confidence, with more mature dogs having an over 10% confidence increase over the youngest. When we compare this to datasets where age was not present, the algorithm averages a 0.6 prediction confidence.

**Analysis 5:** Analysis of trends in likes and retweets for WeRateDogs

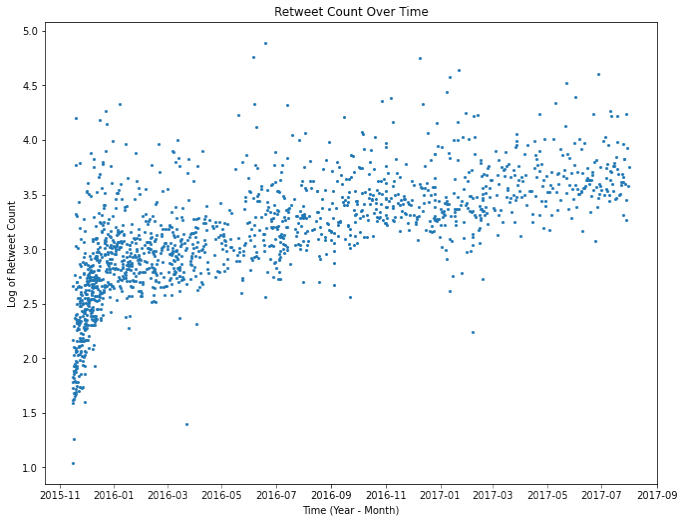


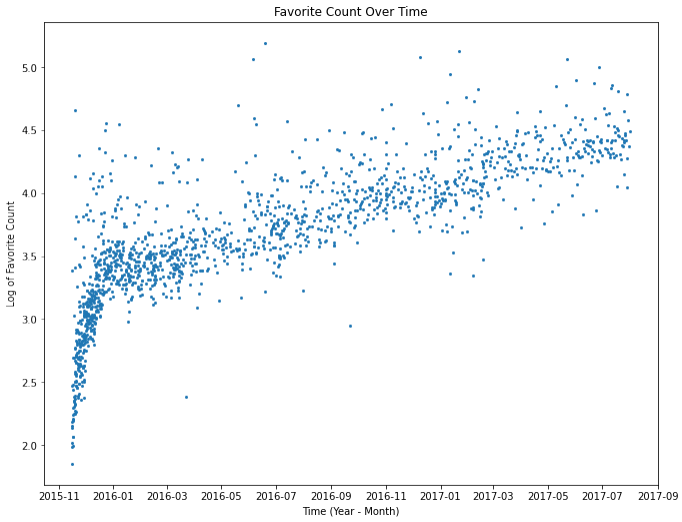




Based on the time span of our dataset grouped by month, there *seems* to be a slightly positive correlation, however the relationship is not strong enough to take for any conclusive analysis. We'll have to change the scale to logarithmic for a clearer picture. First though, let's also look at the relationship between number of likes vs. retweets.

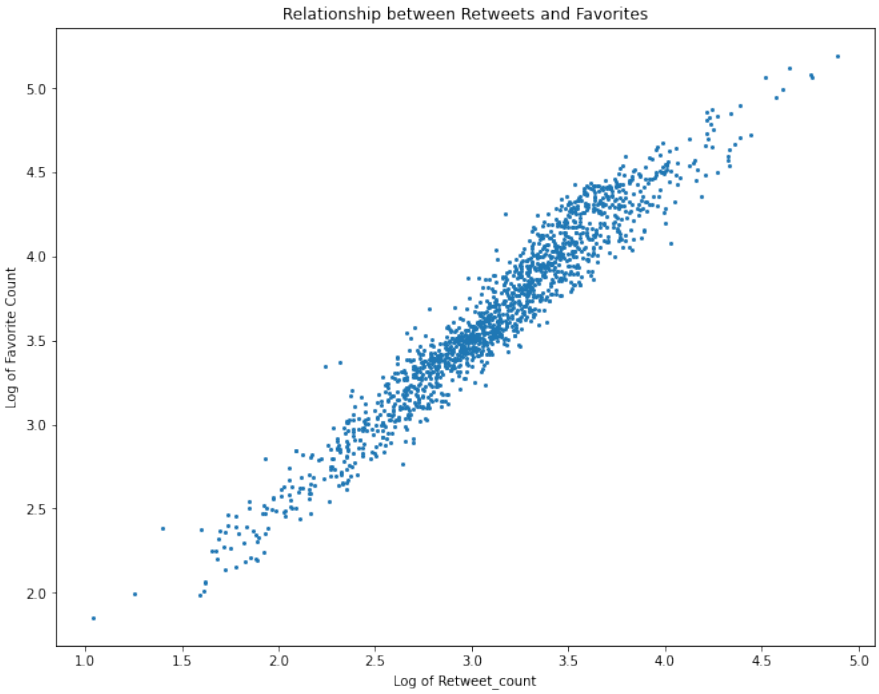



Similarly, the trend *appears* positive, but let's change our scaling to be more confident.




The relationship is much clearer now. We can tell from these charts that WRD experienced a major spike in popularity between Nov. 2015 to Jan. 2016, before tapering out for a more steadied growth. Similarly for the relationship between likes and retweets:




the linear relationship indicates that typically higher liked tweets will also receive more retweets and vice versa, this is to be expected.