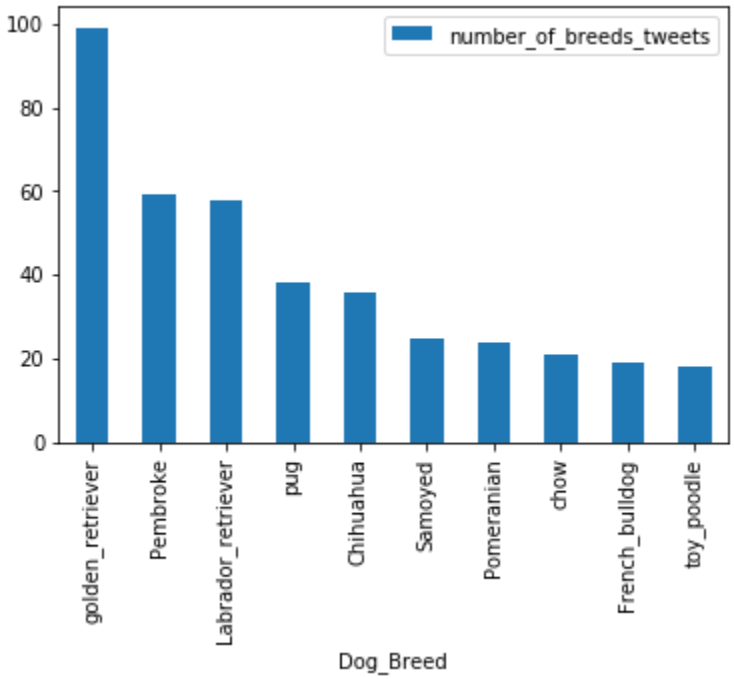
I made three different visualizations:

Fig1:

In this visualization, the x axis is the top 10 of the most frequently tweeted dog breeds, and the number of tweets of each of the breeds.

In order to do this visualization, I merged the two data frames: image_predictions.tsv, and the tweepy object I created with the authorization tokens.

To get accurate data, I limited the p1_conf (the first prediction confidence) to 65% or greater, and removed all the p1_dog == False data (first prediction on if the prediction was a breed of dog or not). I considered these bad data because after checking the p2_conf values of these data points, I discovered that almost all of the values of the p1_dog == False data's p2_conf had less than 1%. To make sure that it wasn't a programmatic error , I opened several of the through the p1_dog == False data's jpg links, and it was evident that almost all of them were either not dogs, or the pictures focus wasn't clear enough to determine an accurate prediction.



Analysis:

There are several noticeable observations to gather from this graph:

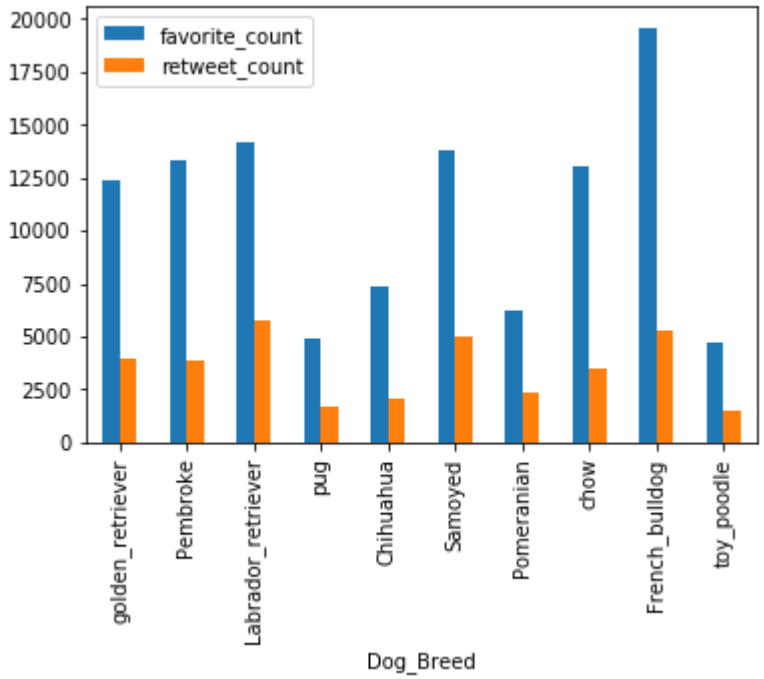
1) The Golden Retriever has a very high likelihood of being the most likely tweeted dog. However, it is possible that the golden retriever might be easier to identify by the image processing algorithm than other breeds of dogs.

2) This is followed by a strong #2 (Pembroke) and #3 (Labrador Retriever). Again, there is a possibility of that these dog breeds might be over-identified due to similar attributes of other dog breeds.

Fig2:

I chose to do this visualization because I was curious as to whether the order of most tweeted dog breeds would be similar to order of their average (mean) likes and favorites respectively.

In order to do this, I made a separate data frame merging image_predictions.tsv, and the tweepy object I created on the dog_breed and gathering each instance of the top 10 breeds and averaging their likes and favorites.



Analysis:

There are several noticeable observations to gather from this graph:

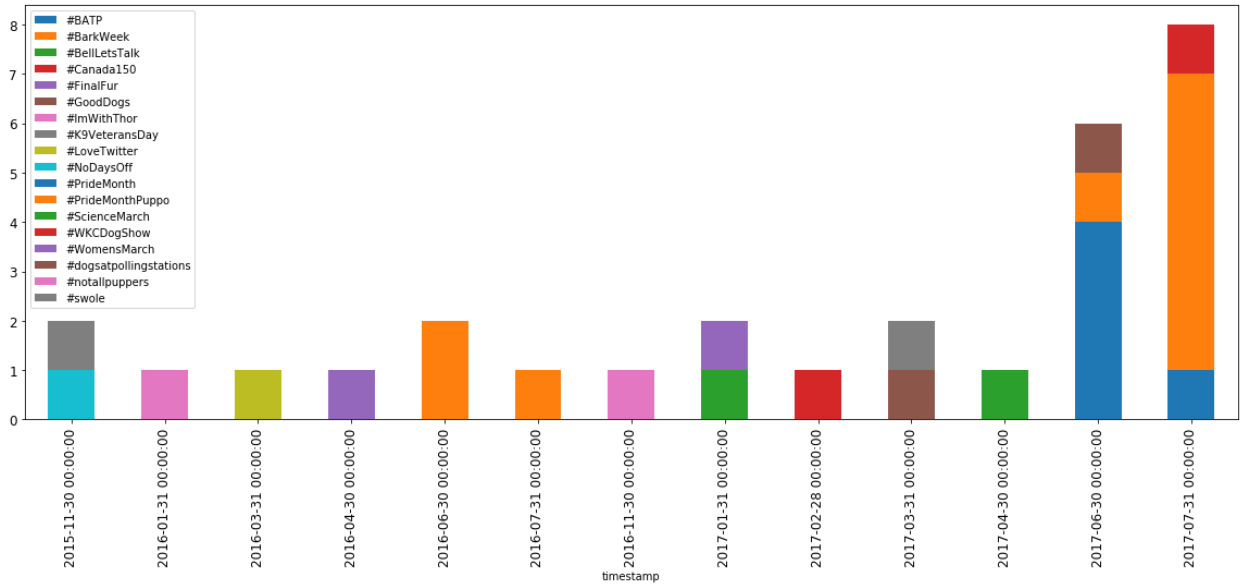
1.) Although the Golden Retriever is the most popular dog breed to tweet, the most favorited breed of dog is by far on average (mean), the French Bulldog. Further statistical analysis could be made to determine possible outliers, but the favorite_count and the retweet_count was gathered programmatically from the tweepy api, and I made sure that duplicates and retweets weren't counted, so I believe that the mean is a good determiner for the average.

2.) The average likes and favorites visually doesn't seem to correlate to the most common dogs tweeted; but the favorite_count and the retweet_count seems to have a likely correlation.

Fig3:

The concept of this visualization was particularly interesting to me; however, the results were disappointing. This visualized the monthly distribution of hashtag instances (ie. #hashtag, #word). I was trying to find if there were any noticeable trends that were specific to certain dates. I assumed that this would be interesting because there could possibly be some future targeted marketing opportunities during certain months of a year due to the past occurrences of hashtags.

In order to do this, I used the enhanced_twitter_archive database, and used a regular expression to grab any instance of a hashtag from each data point's text. Since the data frame covers a period of 13 months, combined all the timestamps into 1 month periods.



Analysis:

This visualization was disappointing because I had assumed that there were going to be more instances of hashtags (there were less than 25 hashtags within over 5000 datapoints). I don't use twitter or Instagram, but I do notice that people use hashtags in text. I thought that this was a programming error on my part, but after reviewing my regular expression and some of the data's texts, it seems that it is not common to use hashtags in either Twitter in general, or on the we-rate-dogs twitter page.

Nevertheless, I did notice some trends that might be worthy of further examination:

1.) The #BarkWeek and the #PrideMonth does seem to happen during the June and July months.

2.) PrideMonth seems to be a newish trend, so that might be something that may be of interest for the branding.

I don’t consider these observations to be significant because there were less than 25 instances of hashtags out of more than 5000 data points.