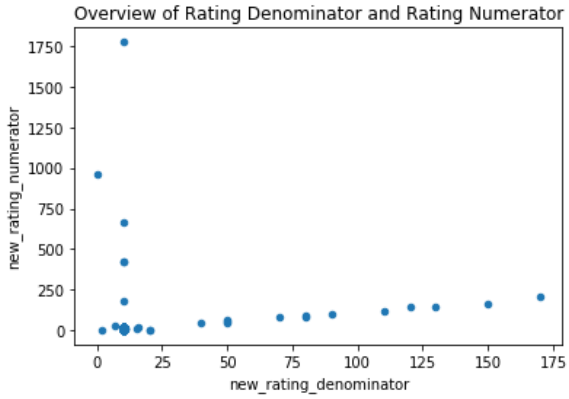
Data Analysis

Limitations

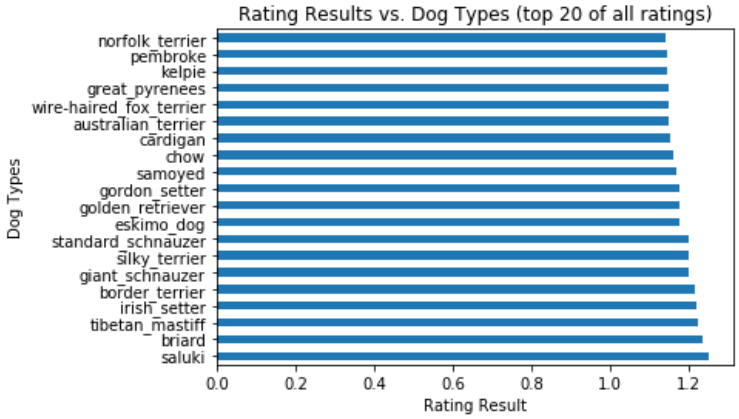
1.**tweet_masterdata**: Unreasonable rating_nominator, although it's probably twitter humor, but there are extreme rating, e.g., 1776, 960, 666, 420. 1451 rows out of 2356 are noted that rating_numerator is higher than rating_denominator. Considering the majority of the rating is unreasonable, no conclusion is made for the relevant analysis. Although the correction according to twitter text has been made, but in general unreasonable rating result is seen.

** ['rating_result'] = ['rating_numerator'] / ['rating_denominator']

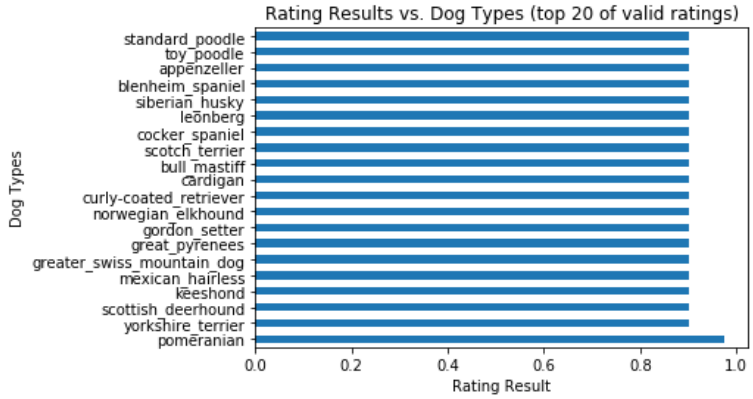
1) The below scatter graph indicates the exceptional cases where numerator and denominators are out of regular rating (10).



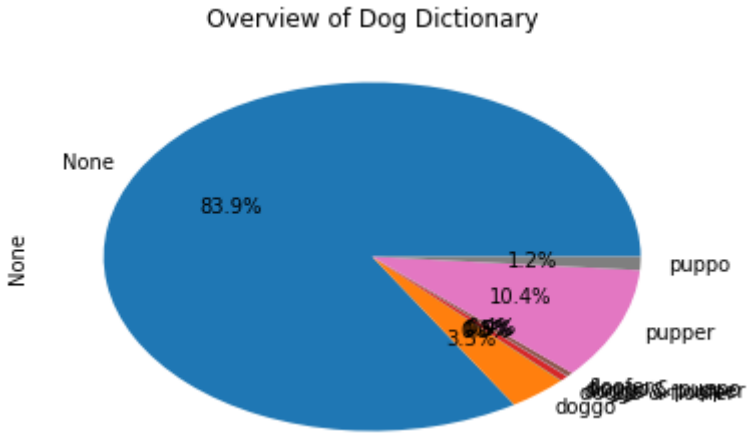
2) The below bar graph shows that the top 20 rating results over 1 (numerator > denominator)



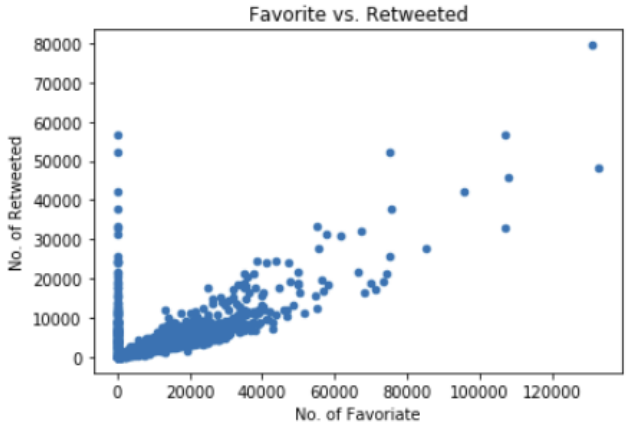
3) After excluding the abnormal rating result (>1), the below bar graph indicates that the top 20 dog types have all most the same rating result (0.8), except "pomeranian". Considering the unreasonable rating denominator, no conclusion is made based on this analysis.



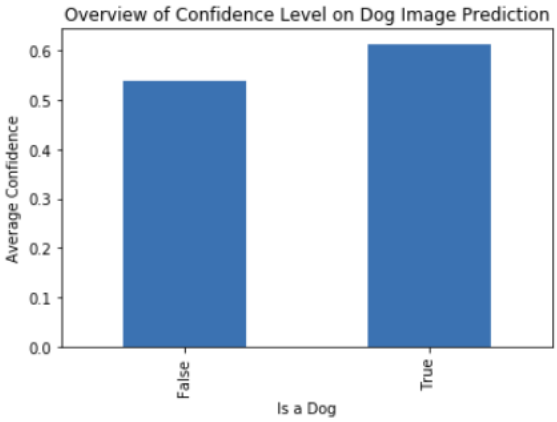
2. **tweet_masterdata**: Although this is information about dog dictionaries (stages), but the majority of the tweets (83.9%) do not have information on this. Based on limited information (16% out of all the available information), we can see 10% is under pupper stage. Moreover, 14 cases noted that there multiple stages are selected, e.g., doggo & puppo. Thus no concrete conlusion can be made based on the limited information.



3.**tweet_overview**: Find 179 out of 2009 valid rows with zero favorite_count, while there is retweet_count, but some posts actually have also non-zero favorite_count at twitter posts.



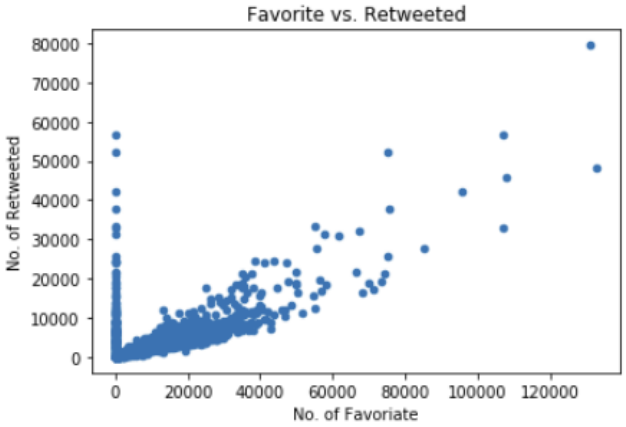
4.**image_predition**: The average confidence of image prediction for dogs is only 61% (non-dog is 54%). Thus it's possible that there are non-dog images but recognized as dogs.



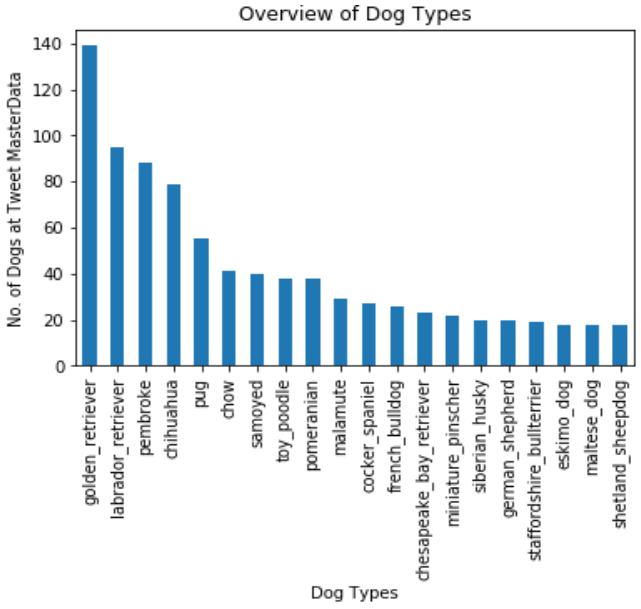
Conclusions:

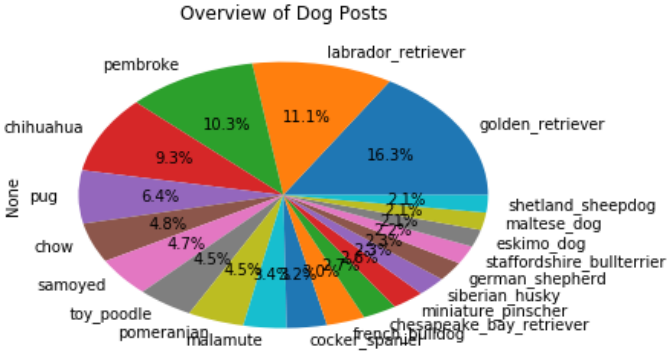
1) Based on valid 1556 non-null data, the top 4 of the most popular dog names are listed below based on numbers of tweet_id:
charlie (12), lucy(11), oliver (11), cooper (11)

2) A positive relationship (70%) is noted betweeen favorite_count and retweet_count which means usually the favorite, usually the post is more retweeted.

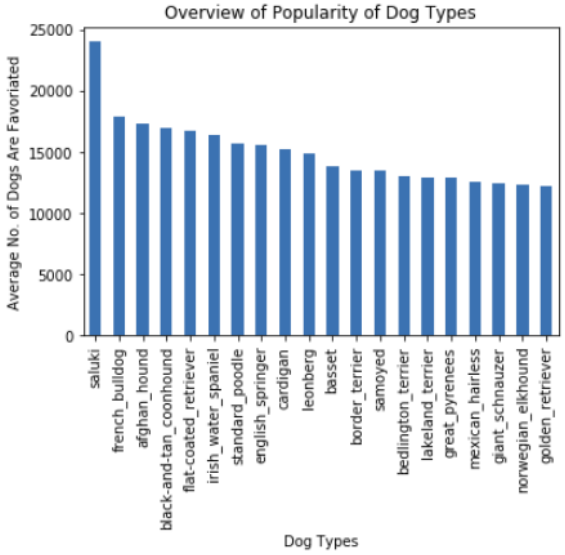


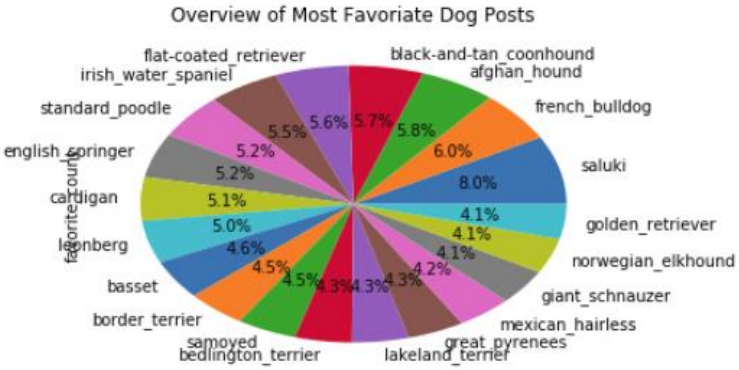
3) Based on the twitter image prediction with valid dog image prediction (1480 rows), golden retriever (16.3%) has the most posts, followed by labrador retriever (11.1%) and pembroke (10.3%), and chihuahua (9.3%).



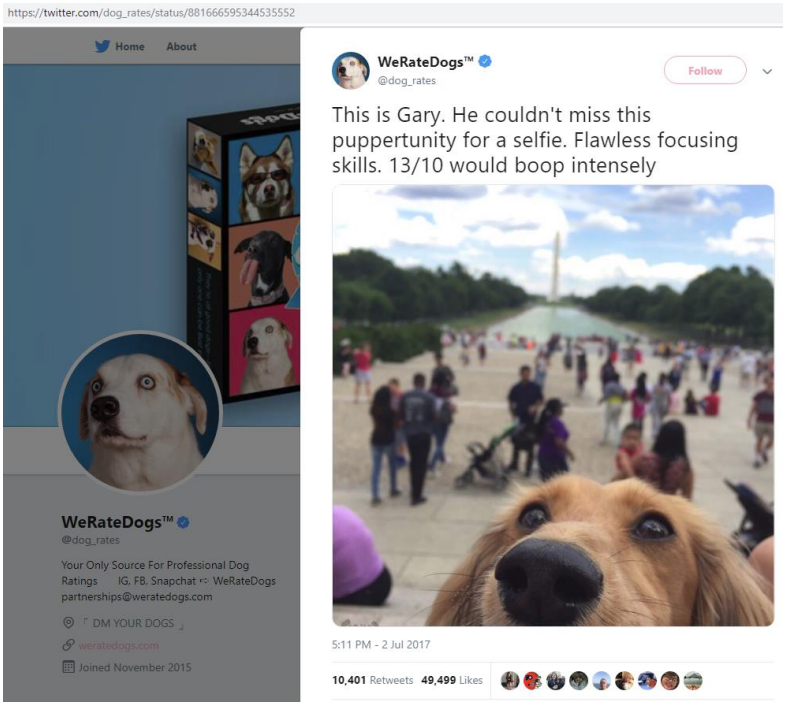


4) It's interesting to note that the most posted dogs don't mean the most popular dogs. For example, based on the top 20 most favorite_count, **saluki (8%)**, **french_bulldog (6%)**, **afghan_hound (5.8%)**, and **black-and-tan_coohound (5.7%)** are the top 4 types of dogs have the most favorite_count.

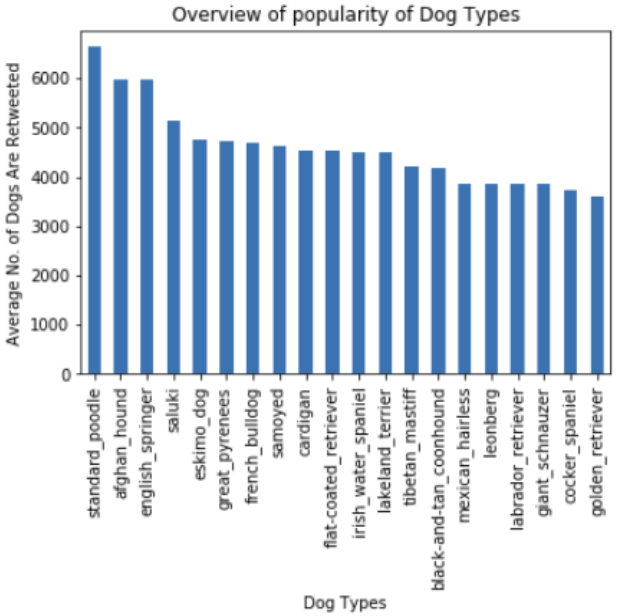


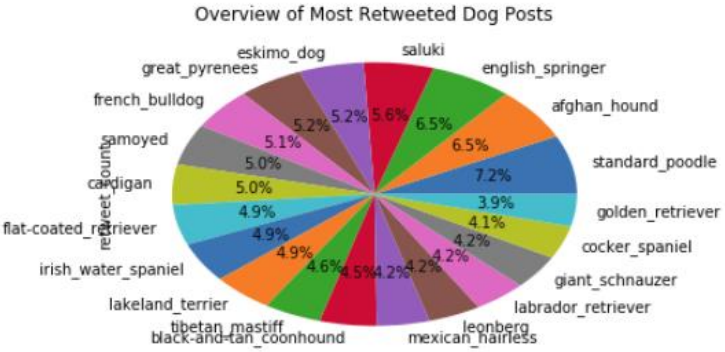


However, while reviewing actual twitter post (example of tweet_id 881666595344535552), the post with most favorite_count seems not because of the dog type, but because of the way how the photo is taken. In this case, it's a dog selfie where we can't really judge the dog type based on the photo and many people reacted also with their dogs' selfie. See below tweet_id 881666595344535552:

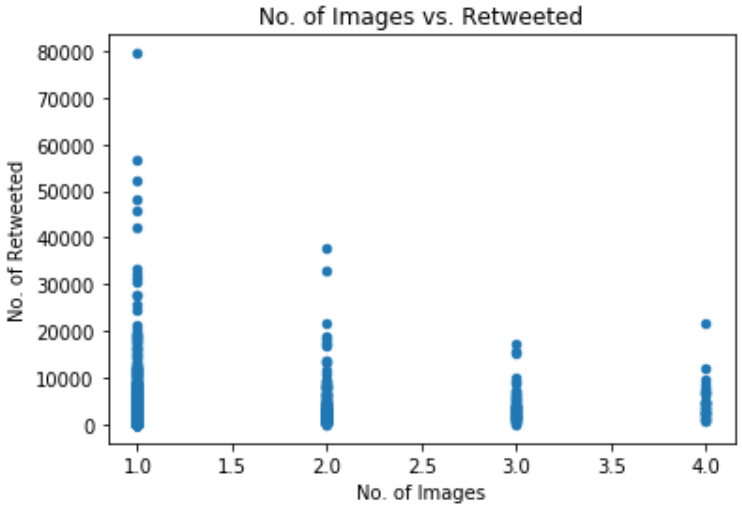


5) Due to the limitation of invalid favorite_count noted (179 rows without favorite_count but do have retweet_count), we also analyze the popularity of dogs based on tweet_count. The result of top 20 most retweet_count is slightly different from top 20 favorite_counts with following: **standard_poodle(7.2%)**, **afghan_hound (6.5%)**, **english_sprinter (6.5%)**, and **saluki(5.6%)**. However, in general, except the top 1 (standard_poodle 7%), the rest top 19 dog types have evenly 4-6.5% popularity among all. Thus we may conclude there are a variety of dog types are liked by people.





6) Moreover, it's interesting to see there is very little relationship (11%) between the number of posted images and popularity. The more posted images don't mean the more retweet_count for example.



7) It's also interesting to see that there is no relationship between rating_results and retweet_count. This is, the higher rating doesn't mean the more popular of the post. However, this is also due to the unreasonableness of rating_numerator.

** ['rating_result'] = ['rating_numerator'] / ['rating_denominator']