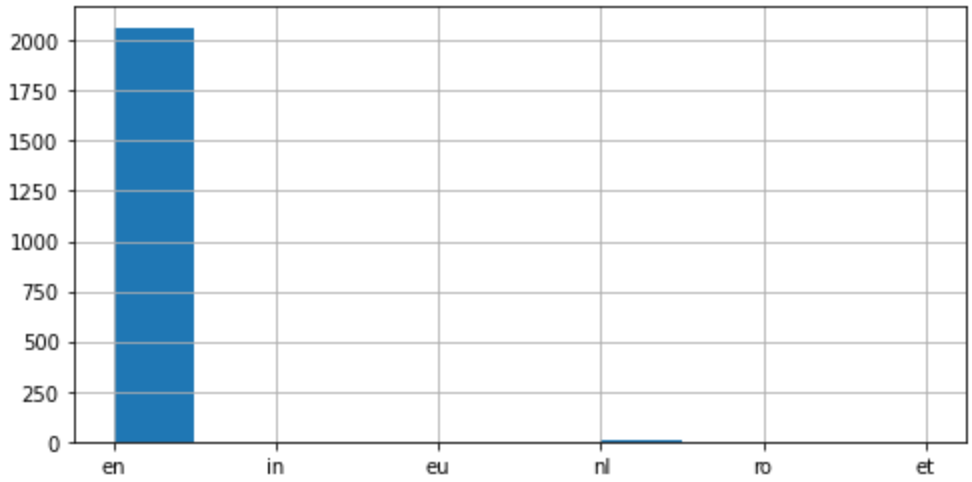
Analysis Report

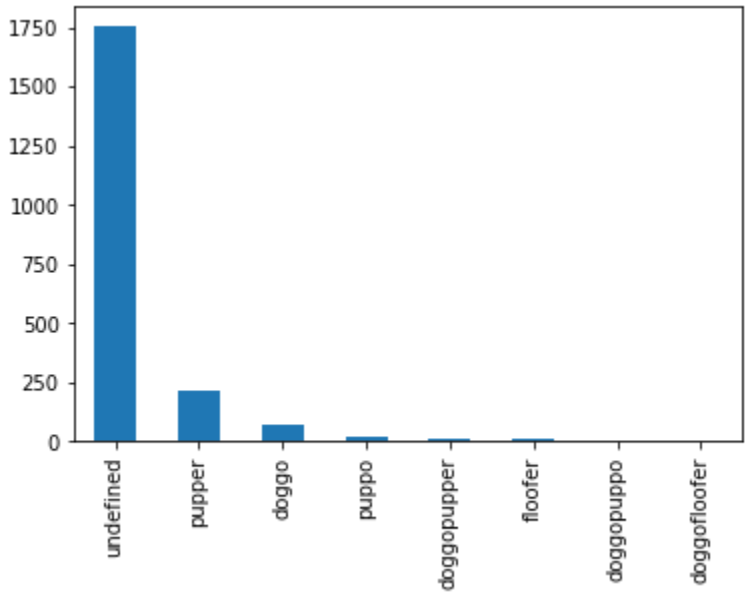
Analysis Conclusion :

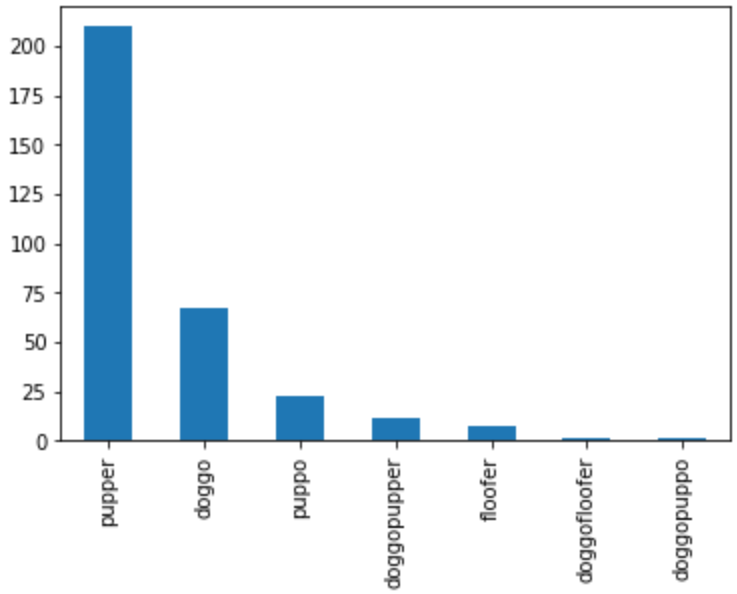
- Charlie, Oliver, Cooper, Penny and Lucy are few of the most common names given to dogs in our dataframe.

- We can clearly infer that, majority of our tweets (close to 99%) are in 'en', which stands for 'English' language.



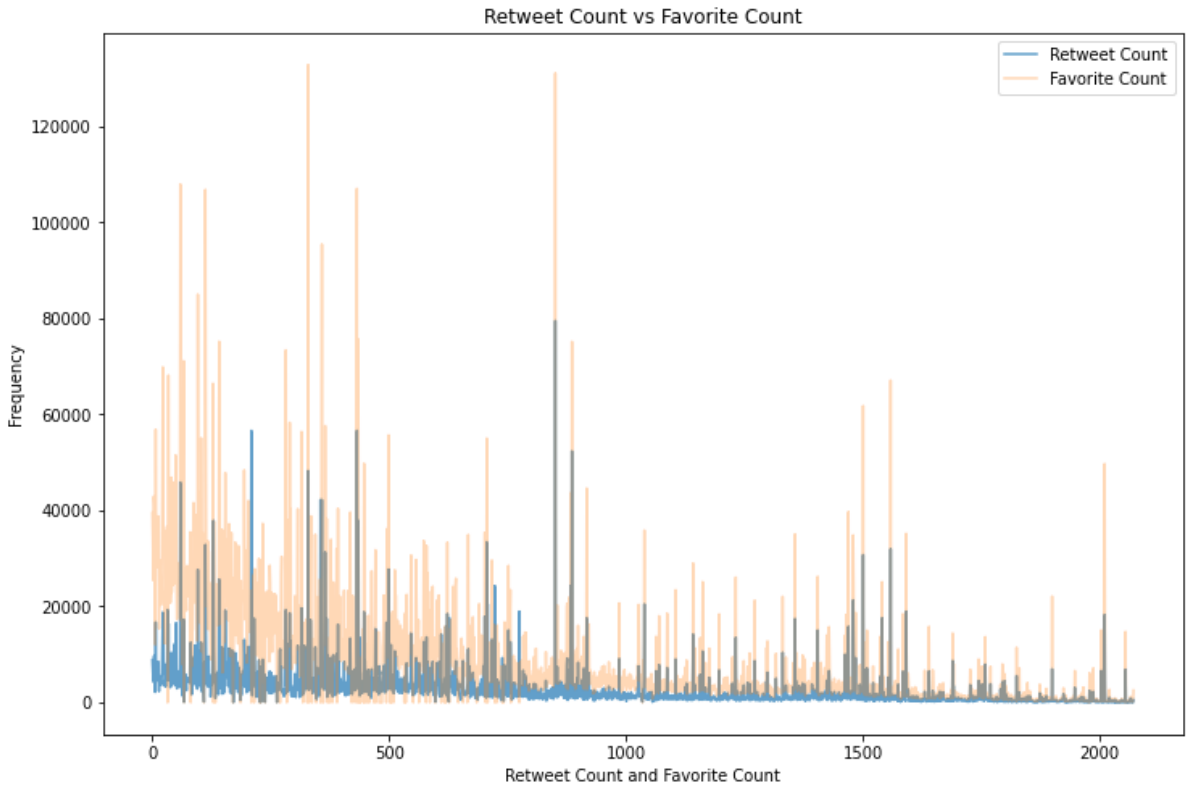
- Most dogs dont have a determined type yet dogs with defined type are : 'pupper'. Followed by : 'doggo' and 'puppo'





- Now Finally after analyzing required last two columns : Favorite Count and Retweet Count we Clear find the following observations:

    - Favorite Count has higher Maximum Value, close to 15000(approx), while Retweet count remains maxed out at 7800 (approx)
    - People tend to add a tweet to their favorite than retweeting a certain tweet.
    - Mean value of Favorite Count > Retweet Count. ('**2976.0892426435116**' and '**8556.718282682103**' respectively)
    - Ratio shared between these two entities (Favorite Count : Retweet Count) = 2.875155139864555: 1 .



Data Quality issues found during Analysis :

1. Data provided by Tweepy API had redundant data, ie; most of its columns has None or NAN values stored and needed to be cleaned before analysis.
2. Data provided regarding DogType is inadequate as most of tweets do not have type mentioned instead, have 'None' values filled.
3. Unnecessary/ Irrelevant data provided in form various columns such as : ['id_str','full_text','truncated','display_text_range','place','contributors','entities','extended_entities'..... etc, which had to be dropped.
4. English being the major language of almost all the tweets provided(99%), lang columns could not be used for analysis purposes.
5. Column Possible Sensitivity has '0.0' as common value for every row throughout the data frame, defying its relevance in the dataframe.
6. The Rating_Numerator column in the data frame has vague values throughout, which in turn had to be engineered from the text column.

7. ID column provided in tweet_json has column name different to other two data frames making alterations compulsory for smooth accessibility throughout the data frames.
8. Data type of Date column is string where it should have been a date and time object.

Tidiness issues found during Analysis :

1. Rows with NULL values existed in all three datasets provided making it untidy.
2. Data scattering and unorganized data found across all dataframes.

Concluding Statement : The conclusions provided do not showcase the complete scenario, as due to reasons such as unavailability of data, we here may have infered biased resuls.