Problem Introduction:

Gathering and analysing tweets data from WeRateDogs tweets.

Efforts:

1. **Data Gathering Efforts**

   i. locally stored twitter-archive-enhanced.csv
   ii. Downloading https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
   iii. Get additional information using Python's Tweepy library e.g. retweet count, likes
   iv. Each tweet's JSON data should be written to its own line.
   Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2. **Quality and Tidyness Issues Identification**

   2.a Qualtity Issues

   i) Coulmns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp having large numbers of NaN

   ii) 2 tweet_ids not found in data obtained through api. 2356 ids are there in twitter-archive-enhanced.csv wheras tweet_json.txt contains 2354 unique tweet ids.

   iii) 281 tweet_ids are missing from image-predictions.tsv in comparison to twitter-archive-enhanced.csv

   iv) None and a do not seem to be names, 576 names are "None" and 55 names are "a"

   v) 5 ratings numerators not extracted correctly in the rating_numerator column

   vi) 1 ranting's denominator not extracted correctly

   vii) data type for the following columns should be changed:

   viii). TimeStamp should be chnaged to datetime

   ix). rating_numerator, rating_denominator type should be changed to float

   x) Source column does not contain full source content and has curtailed source data which is not useful

   2.b Tidyness issues:

   i) Dog stages has 4 columns which can be consolidated in to one column.

   ii) tweets info are distributed in different tables which can be merged in to one.

3. **Conclusions:**

   a. Higher retweets were seen for higher rated Dogs, as expected

b. Popular names are checked and based on the numbers top 5 popular names were Charlie, Penny, Cooper, Tucker, Lucy and Oliver

c. By far golden retriever looks to be most popular breed which has got recognized with more than 50% probability