Gathering Data

We have three sources of data in total. We have the first and second dataframe by accessing twitter_archive_enhanced.csv and image_predictions.tsv.
Then , using tweet_id from twitter_archive_enhanced.csv and Tweepy library, we get the retweet amount and like amount for each tweet as the third dataframe.

Assessing Data

First, we check duplicates using code.
Since each tweet has a unique id, we check uniqueness on the tweet_id column.
We have no duplicate tweet in all of the dataframe.
Second we check null using code.
We notice that in df_1, we have null values. However, every tweet has a rating and dog's name, which means we do have a dog with a rating for each row. The null values are only in the not important column. Therefore we do not drop any rows.
Then we check values in general.
By checking minimum and maximum, we notice either we have outliers or we have some data incorrect.
By checking uniqueness, we notice the non-null jpg_url are 2075 but we only have 2009 of them being unique. It could be retweeted if the image is repeated.
Finally we read the data roughly by google spreadsheet.
We notice that, for the outliers on rating, there might be something wrong.

1. Some rows are rated for goats and cats. We should delete them since they are not rated for dogs.
2. We have ratings with wrong numbers, we can correct them by reading from the original tweet text.
3. Some names are not correct. Some dogs' names are misunderstood by random words. If a dog has no name, we should change the wrong name to None. Notice that in this case, the misunderstood words start with lower cases characters. We can change them by string search.
4. Retweets are not what we are interested in, we should delete them.
5. Columns 'doggo', 'floofer', 'pupper', 'puppo' are values instead of column names. There are rows that have multiple dog stages.
6. We do not have the same tweet amount in all of 3 dataframes.
7. We should check the data type for numbers and change them if necessary.

8. We might want to add a column for rating_numerator / rating_denominator since there are 18 unique rating_denominator.

Cleaning Data

1. Change df_3 data type to integer.
2. Delete retweet tweet, deleting rows by searching whether the string is started with 'RT'.
3. Delete df_1 tweets that are not dogs.
4. Change df_1 column rating_numerator data type from integer to float.
5. Change df_1 incorrect rating error.
6. Delete df_1 incorrect rating row 518.
7. Change df_1 incorrect naming.
8. Change dogs' name to 'None' by searching whether the first letter is capital or not in column "name".
9. Add a column called rating by rating_numerator / rating_denominator.
10. Merge doggo, floofer, pupper and puppo to one column called dog_stage and change multiple stages.
11. Use inner join to merge three dataframes with tweet_id.