## Wrangle Effort Report

1. When I use .info() function to df1, I noticed its "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", and "retweeted_status_user_id" are floats, which should be integers in the real world. Therefore, I used the .astype(np.int64) function changed these columns into integers.

2. There were many null values in the "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", and "retweeted_status_user_id" columns. Besides that, I thought df1 table contains too many information, meaning each type of observational unit didn't form a table. This redundancy of data was a reason that there were too many NaN values in the columns. Therefore, I decided to separate df1 into three tables, one called "retweet" stores info about retweets, one called "reply" stores info about the replies, the last one is the table that contains the rest columns.

3. Still by using the .info() function, I notice the data type of "timestamp" and "retweeted_status_timestamp" are strings, which should be datetime type in the real world. Therefore, I used the pd.to_datetime function converted their data types into datetime.

4. The data I actually wanted were tweets info of non-retweets that have images. I first merge df1_clean and df2_clean with inner method so that I can get the tweets that are both in df1_clean and df2_clean. (To make this part cleaner, I first created a new column in df2_clean called breed, which I concluded from df2's test result columns, and then merge the breed column in df2_clean with df1_clean on the "tweet_id" key.) Then I dropped all the rows that were retweets of the after-merged df1_clean table.

5. I found the strings in breed columns were sometimes in capital-case and sometimes were in lower-case. Therefore, I lowered all the strings in the breed column.

6. I noticed some of the rating were wrong, the original rating info were the first match when extracting the text column. But actually, the exact rating is not the first one. And sometimes there are two dogs and two ratings in one tweet. So, I extracted all matches from the text column and checked each row that had over one match. Gave average rating to tweets had over one rating, and fixed the wrong ratings.

7. I noticed there was a tweet (index=137) incorrectly extract "puppo" from "puppon" from the text, which was wrong, the stage should be "doggo". And some of the rows has more than one stage as well. So, I extracted all the matches in a list from the text column. Added the first matches in this list to df1_clean, named the new column "stage". And then I checked each row that had over one match, fixed the wrong stages, and gave a new stage name to the rows had over one stages. Fortunately, all rows that had over one stage had only doggo and pupper stages at the same time, so I gave them a new stage name "doggo_pupper".

8. When I was dealing with the first research question, I created a new column in df1_clean called "rating" rating", which was calculated by using rating_numerator divided by rating_denominator. Then I realized that some rows had extremely high rating, so I went back to the data cleaning step and check the rows that had really high ratings, fixed the wrong rating and drop the rows whose ratings were too high to make sense.