## REPORT OF WRANGLING OF WERATEDOGS TWITTER ARCHIVE

The data was messy and untidy, I wrangled it and cleaned the most obvious of the mess and untidiness. I cannot say that it is 100% clean, but I can say is it is clean on the average. I cleaned 2 structural mess and 11 quality mess.

## UNTIDINESS CLEANED

1. Making df_twitter_archive to have equal number of rows with df_new.

2. Making `timestamp` in df_twitter_archive to become `tweet_timestamp`.

3. Creating a retweeted table from df_twitter_archive.

4. Merging the dataframes as they are of the same observational unit.

5. Melting df_twitter_archive columns - doggo, floofer, pupper and puppo to form rows instead and remove resulting duplicates.

## QUALITY MESS CLEANED

1. Making `in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id` to become int.

2. Making `retweeted_status_timestamp and timestamp` to become datetime and not string.

3. Making all `rating_numerator` not greater than 10 to become the mode of the rating numerator field.

4. Making all `rating_denominator` to become exactly 10.

5. Changing all names of dogs bearing 'a' or 'an' or 'the' or more generally starting with lower case to be np.nan.

6. Droping tweets that threw HTTP Error and Tweepy Error from df_twitter_archive.

7. Removing data duplicates after the melting of columns `doggo, floofer, pupper, puppo`.

8. Removing all underscores in `tp1, tp2, tp3` entry.

9. Droping `in_reply_to_status_id,in_reply_to_user_id` since it all contains null values in  retweets table created from df_twitter_archive and drop all `in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id` in the twitter_archive.