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

- I separated the tweets that were retweets from original tweets and kept them aside in their dataframe ‘retweets’
- The structure of the ’retweets’ and ` twitter-archive-enhanced’ file were changed to obey the rule of tidy data. I melted the doggo, floofer, pupper and puppo columns to become to one column ‘growth_stage’, and their original values were stored in a field `growth_stage_value’

QUALITY MESS CLEANED

- First the ‘twitter-archive-enhanced’ file was accessed and tweet_ids that threw HTTP and Tweepy exceptions were dropped from the file or dataframe.

- I made `tp1_confidence, tp2_confidence, tp3_confidence` datatype to become float since they depict the certainty of prediction which is a real number.

- I made `img_num, in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id` to become int.

- I made `retweeted_status_timestamp and timestamp` to become datetime and not object.

- I made `tp1_dog, tp2_dog, tp3_dog` field to become bool.

- I made `tp1_confidence, tp2_confidence and tp3_confidence` float values to have equal number of precision at least 7 decimal places.

- I made all `rating_numerator` not greater than 10 to become the mode of the rating numerator field.
- I made all `rating_denominator` to become exactly 10
- I changed all names of dogs bearing 'a' or 'an' or 'the' to be None
- I made `in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id` null values to become 0 instead of Nan
- I made `timestamp` in df_twitter_archive to become `tweet_timestamp`
- I went further by cleaning the retweets dataframe making the datatypes of the columns to conform to standard.