Reporting: Wrangle Report

Issue 1: Join `doggo`, `floofer`, `puppo` and `pupper` columns into one single column

The values for these columns are the same as the name of the column itself. The columns should be combined into a single column and it will represent classification of dogs.

Issue 2: `df` dataframe contains data beyond 1st August of 2017, for which there is no image prediction data in `df_img` dataframe

For image recognition there is no data beyond 1st of August 2017. Because of this we need to change format of timestamp to datetime, filter and drop rows after above mentioned date.

Issue 3: `df` dataframe includes retweets

Data contains retweets, which affects consistency of our data. Retweets can be identified through retweet_status_id column. We need to drop rows where there is data available under above mentioned column.

Issue 4: In `df` dataframe need to remove rows, where values are not Null under `in_reply_to_status_id`, because they don't provide actual rating

`in_reply_to_status_id` seem to be just replies to someone posts and creates quality issues for our data. The same way, as was done above, rows which contain data under `in_reply_to_status_id` need to be dropped.

Issue 5: In `df_img` not all tweets are regarding dogs, need to leave tweets regarding to dogs and drop the rest from both `df_img` and `df` dataframes

`df_img` dataframe provides 3 options, which describes breed, confidence coefficient and whether its dog or not. Firstly, we need to retrieve the tweet_ids of those that are not dogs to a list. Later this list will be used to drop the same tweets from `df` dataframe. The tweet ids will be selected where in all three cases the prediction is `False`. In other cases, if the one with highest confidence coefficient has `False` as dog prediction.

Issue 6: `tweet_id` column in `df`, `df_img` and `tweep_json` dataframe is integer instead of string

Later tweet ids will be used for merging dataframes. So their formats need to be changed to string to avoid errors.

Issue 7: Merge `retweet_count` and `favorite_count` columns from `tweepy_json` and breed from `df_img` to `df` dataframe

Merging these three dataframes will solve tidiness issues. The dataframes will be merged to tweet ids. In `df_img` case, a list of dictionaries will be made, which holds tweet_id, breed and being a dog Boolean (which will be used to avoid errors at first and then dropped in later stages). This list will be used to append data to `df` dataframe. In case of `tweepy_json`, we will need just `retweet_count` and `favorite_count` columns and these columns will be merged with `df` dataframe.

Issue 8: Drop redundant columns in `df` dataframe and missing data

Columns that helped us to find retweets and replies will be dropped. Also being a dog Boolean. We also won't need doggo, floofer, pupper and puppo columns, since they are not combined. Name column gives no informational value and there is a lot of misleading and null data, but for the tweet ids, breed and other data can be identified. So, name column should be dropped. Other than these, we will use pandas.dropna() to clear out any missing data.

Issue 9: `rating_numerator` doesn't match with the rating provided in the text

Visual analysis of csv file revealed that in some cases rating numerator doesn’t match the rating in the text. Such rating were floats. Using pattern "(\d+(\.\d+)?\/\d+)", the ratings were extracted from the text and replaced rating_numerator using these. Since the extracted data was string, their format was changed to float.

Issue 10: `rating_numerator` and `rating_denominator` should be expressed in single column in `df` dataframe

Visual and programmatic analysis revealed, that in some cases rating_denominators were above 10 (which in most cases was 10). Analysis revealed that the denominator was higher, when number of dogs in a picture was more than 1. Rating_denominator = number of dogs x 10. For this reason, it will make farther analysis of the data easier, if we will have single rating column. The best option will be division of rating_numerator by rating_denominator.

Issue 11: In `df` under `source` column, links have html tags and they should be removed

Source column contains links to twitter application. It will be much more useful if the column contains only the name of the device used for tweeting. We can see with value_counts() that there is only iphone, vine, web client and tweetdeck used for tweeting. So we will identify these texts in text and replace them with just the items from the list above.