Wrangling Efforts

Three dataset: “twitter_archive_enhanced.csv”, “image_predictions.tsv” and “twitter_api.py”. The “twitter_api.py” is downloaded as “tweet_json.txt”, but it is actually a JSON file. After **gathering** all the date, I read them in Jupyter Notebook. Then I **assessed** the data visually and programmatically to find quality and tidiness issues. Assessing data helps to observe the first few and last few rows and also arbitrary rows of whole dataset. Assessing helps to find the quality issues such as: number of rows and columns, nun values, type of values, number of unique values in each column, number of duplicated rows, general statistics (mean, standard deviation, min, max, first quarter (25%), second quarter (50%), and third quarter (75%), name of columns, and wrong data. Besides quality issues, assessing also gives information about tidiness issues such as: columns with similar information in different columns, unnecessary combined columns. The observed issues in the provided dataset are:

Quality issues

1. In twitter_archive, image_predictions, and tweet_json dataframes all id columns (such as tweet_id) should be string.
2. In twitter_archive the timestamp column should not be object. It should be date type.
3. Name of id column in tweet_json is different from the other two dataframes (tweet_id)
4. In twitter_archive dataframe, the name of 109 dogs are wrong. By mistake their name is: a, an, the, quite...
5. In twitter_archive dataframe, the name of 745 dogs is "None" which is wrong.
6. In twitter_archive dataframe, the name of dog for twieet_id of "885518971528720385" should be Howard
7. In twitter_archive dataframe, in the text file there are many compains about "presenting the pictures which does not contain dog picture such as:

    883117836046086144 Please only send dogs. We don't rate mechanics...
    880872448815771648 Ugh not again. We only rate dogs. Please don't...
    880095782870896641 Please don't send in photos without dogs in th.
    872486979161796608 We. Only. Rate. Dogs. Do not send in other thi...

8. In twitter_archive dataframe, sometimes the rating_nominator and rating_denominator of text column is different from the current rating_nominator and rating_denominator. Example: Row 45 the rating_nominator and rating_denominator from text files are 13.5 and 10, however in the rating_nominator and rating_denominator columns it is 13 and 10!
9. In twitter_archive dataframe, the rating_denominator of 23 record is not equal to 10 and for every other record is 10. It seems incorrect.

Tidiness issues

1. In twitter_archive, there is no need for having 4 coulumns presning different dog stages
2. In twitter_archive, in timestamp column, date and time are combined. They should be in two different columns
3. In image_predictions, there are three model for predicting breed. We only need one with highest confidenc einterval.

4. We do not need three dataframe for these project. One dataframe should be enough.

After finding the quality and tidiness issues, I **cleaned** the data in three steps: 1) defining the issue, 2) coding to fix the issues, 3) testing if the issue is fixed.