Data wrangling

Gather the data

The data wrangled in this project corresponds to the WeRateDogs Twitter data. Three different sources of data (csv, json and tsv) are loaded in pandas dataframes for further analysis. The csv data includes the enhanced Twitter archive which contains tweet information improved with the dogs “stage”. The json file contains additional raw data, tweets extracted from twitter. Finally the tsv file contains image prediction on the object or type of dog present on the picture.

Assess the data

The **enhanced Twitter archive** file presents some quality issues:

- "None" values in the last columns are not captured as missing data
- the rating_denominator is showing values greater than 10
- the column with dogs' names presents 55 times the name 'a' which seems to be a mistake

And tidiness issues as well:

- the variables: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id" and "retweeted_status_user_id" are showing a float type instead of integer
- the variables: "timestamp" and "retweeted_status_timestamp" show an object type instead of datetime

The following observations were made on the quality of the **raw data from json file**:

- the variables: "geo", "coordinate" and "contributors" do not contain any value other than nan
- the variable "place" contains only one non-missing value
- the variables "retweeted_status" and "quoted_status" contain 179, and respectively 28, non-missing value. Each one being a dictionary with the fields in dataframe tweet_json from column "created_at" to "lang"

And its tidiness:

- the variables: "in_reply_to_status_id", "in_reply_to_user_id", "in_reply_to_status_id_str", "in_reply_to_user_id_str", "quoted_status_id" and "quoted_status_id_str" are showing a float type instead of integer

Finally the dataset from **tsv file with image predictions** looked clean and tidy. There is no missing data.

Clean the data

A copy of the data to new dataframes was made before cleaning.

In the **enhanced Twitter archive,** the None values were replaced with nan values.
The rows for which the rating_denominator is greater than 10 were removed.
The dogs' names with a value 'a' were replaced with nan values assuming that those names were due to mistakes while typing.
For the following variables, their type was changed to integer: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id" and "retweeted_status_user_id".
For the following variables, their type was changed to datetime: "timestamp" and "retweeted_status_timestamp".

In the raw data from the json file, the columns "geo", "coordinates" and "contributors" were dropped since they do not contain any value other than nan.
The column "place" was dropped as well, since it contains only one non-missing value.
The columns "retweeted_status" and "quoted_status" were also dropped as they present less than 10% of non-missing data.
For the following variables, their type was changed to integer: "in_reply_to_status_id", "in_reply_to_user_id", "in_reply_to_status_id_str", "in_reply_to_user_id_str", "quoted_status_id" and "quoted_status_id_str".

Finally, the three datasets were merged into one dataset using the tweet id, which has been saved as a csv file titled “master_dataset.csv”.