Gathering Data

First wrangling step is gathering data. Here we will work on 3 datasets

1. "twitter-archive-enhanced.csv" after importing this file we changed it to a data frame to work with it programmatically.
2. “image-predictions.tsv” the second dataset that we have. First we download it from a given URL programmatically using requests methods.
3. “tweet-json.txt” this file has all tweets details from it we can get a good data for analysis after importing it we changed it to a data frame to make it easier to deal with.

Assessing Data

The second step is assessing data to make it good and arranged dataset by investigating it carefully to know how to make it a usable and good dataset after doing some fixes:

Quality issues

- **Completeness**
  - missing(NaN) values in columns (in_reply_to_status_id,in_reply_to_user_id) and breed types
  - missing values in name column that can get it from text
  - tweets with no images
  - datasets have retweets and replys data not only tweets

- **Validity**
  - Not valid names (the, a, an, by)
  - There are tweets with out images (expanded_urls and retweet url are empty)
  - There are retweets in all datasets in this data (
    - retweeted_status_id 181 non-null float64
    - retweeted_status_user_id 181 non-null float64
    - retweeted_status_timestamp 181 non-null object)

  - there are replys in all datasets in this data (
    - in_reply_to_status_id 78 non-null float64
    - in_reply_to_user_id 78 non-null float64)
  - rating_numenator col doesn't match with rate in text (ex: index 47) and there is outliers
  - denominator not equal 10 (ex: index 435)

- **Accuracy**
  - name values don't match with names in text
  - rating_numenator col doesn't match with rate in text (ex: index 47) and there is outliers
  - source column is an HTML tag

- **Consistency**
  - tweet_id dtype is int
  - unused columns like (denominator, expanded_urls, retweeted_status_id, retweeted_status_user_id, in_reply_to_status_id, in_reply_to_user_id)
  - unused columns like img_num and p_n_prediction in image_p df

Tidiness issues

- values are column names in archive_df (ex: doggo, floofer, pupper, puppo should be combined into a single column as this is one variable that identify type of dog)
- values are column names in image_p_df (ex: p1, p1_dog, p1_conf. p2, p2_dog,p2_conf,p3,p3_dog,p3_conf all these columns should be in only three columns)
- Information about one type of observational unit (tweets) is spread across three different DataFrames. So these three DataFrames should be merged as they are part of the same observational unit.

Cleaning Data

After that we know the issues then we go to the third step which is cleaning data. Here we will define all cleaning issues.

Define

- remove retweets, replys and expanded_urls from all datasets
- drop columns:
  - expanded_urls,
  - in_reply_to_status_id,
  - in_reply_to_user_id,
  - retweeted_status_timestamp,
  - retweeted_status_user_id,
  - retweeted_status_id

- fix breeds' name by making one new col and drop remain 4 cols
- change name col values
    - - get untitled names in clean_archive df then assign that name to "NaN"
    - - get lower names in clean_archive df then assign that name to "NaN"

- extract name from text from clean_archive_df
- extract rating_numerator and rating_denominator if a rating_denominator is more than 10
- deal with rate
- fix source col in clean_archive_df
- rename cols in clean_image_p_df
- get breed name and confidence from clean_image_p df to avoid values in headers issue
- drop unimportant cols in clean_image_p df and merge tmp df instead
    - "prediction_1", "confidence_1", "breed_1","prediction_2", "confidence_2", "breed_2", "prediction_3",\"confidence_3", "breed_3"
- merge clean_image_p and clean_json and clean_archive df in twitter_archive_master df
- drop the row that has sentence "We only rate dogs" this tweet's image doesn't include dogs
- replace None values to Nan in type and name columns
- drop rating_denominator col from twitter_archive_master it's a known fixed number
- drop img_num col from twitter_archive_master not needed in analysis
- remove a row has no name&type&breed & numerator >100 (ex: twitter_archive_master.rating_numerator == 420.0 not a dog)
- change tweet id dtype to string
- reset index in twitter_archive_master df