# Data Wrangling Process

## Data Gathering

Download data file that contain dog breed predictions based on tweet image, and save to file: image\_predictions.tsv.

Use twitter\_archive\_enhanced.csv as the center piece of data, extract its tweet\_id’s. For each tweet\_id, download its json data from Twitter website archive, save to tweet\_json.txt.

## Data Combining

Load twitter\_archive\_enhanced.csv as the main data frame. Merge with the dog breed predictions file by tweet\_id.

Load tweet\_json.txt to data frame, select only id, retweet\_count, favorite\_count, and merge the counts into the main data frame by (tweet) id. In tweet\_json.txt data, id and id\_str don’t always match. But since id\_str is not used in later analysis, the issue does not need to be fixed.

Save the combined data frame (main data frame) to an intermediate data file: combine\_tweet\_data.csv

## Data Cleaning

* Retweets: we do not want to keep retweets, so I remove rows that has non-null retweeted\_status\_id, and then drop columns: retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp
* Rating\_denominator: there are 18 entries with various values other than 10, remove these rows from the main data frame.
* Rating\_numerator: there are 5 entries with values > 20, while the rest are <= 14. Remove these 5 rows.
* In\_reply\_to\_user\_id: this information is irrelevant, drop columns: in\_reply\_to\_status\_id, in\_reply\_to\_user\_id

Data frame df is saved to df\_1.csv

* Combine dog type column doggo, puppo, pupper and floofer into a single column. In the consolidated column, 11 tweets have duplicated dog types: 'doggo' and one other type. Since 'doggo' is a more generic name, remove these duplicate rows with 'doggo' type.
* Dog names: a sorted unique values list of dog names show that there are names that seem to be erroneous: 'a','actually', 'all', 'an', 'by', 'getting', 'his', 'incredibly', 'infuriating', 'just', 'light', 'my', 'not', 'officially', 'one', 'quite', 'space', 'such', 'the', 'unacceptable', 'very'. Replace these names with ‘None’.
* Since rating\_denominator is always 0, remove this column

Data frame df is saved to df\_2.csv

* P1\_dog impact on p1: for image prediction of dog breed that is not a dog, set p1 to Nan ; remove p1\_dog column as it is redundant now
* P1\_conf impact on p1: very low confidence (<15%) in dog breed prediction is not reliable, set those predictions p1 to NaN.

Save the final cleaned data frame to tweeter\_archive\_master.csv.