# Data Wrangling – Assessing, cleaning, and storing the data

## In this section below activities are performed and explained:

- Gathering
- Assessing
- Cleaning
- Store

### Gathering

The data has been gathered in three data frames from three different sources:

1. archive_df  from 'twitter-archive-enhanced.csv'
2. image_df  from 'image_predictions.tsv'
3. tweet_json  from 'tweet_json.txt'. This file contains fields required for analysis.

***Note***:

· I have followed the directions for accessing the Twitter data without actually creating a Twitter account

· All the files are attached to the project submission zip.

### Assessing and Cleaning

I have performed an analysis of all the three data sets and below are the Data Quality and Tidiness recommendations:

The data has been cleaned up as per the recommendations.

### Twitter Archive Data

1. Missing data for columns in archive_df - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,

retweeted_status_user_id, retweeted_status_timestamp. It could be due to retweets, which can be deleted.

2. Missing data for columns in archive_df - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp. 2. There are names such as 'None', 'a', 'such','O','a', 'actually', 'all', 'an', 'the', 'this', 'unacceptable', 'very','my', 'not', 'officially', 'by', 'getting', 'his'.
3. Most of the names start with Caps while there are few at the end starting with small letters.
4. The four columns, doggo, floofer, pupper, and puppo has all the values populated as None for 1976/2356 records, so there is no dog "stage" (i.e. doggo, floofer, pupper, and Puppo) information.
5. The rating numerator is 1776 for one record and 0 for two records.
6. The rating denominator is expected to be 10 but there is another value too.
7. We are not interested in retweets and there are 181 retweets in the archive data.

### Image prediction data

8. There are 324 records for which all three algorithm predictions is other than the dog breed. We could save time by removing these records.
9. The dog breed names populated in the p1, p2, and p2 are not consistent few are starting with CAPS while others are with lowercase.
10. There are 66 duplicate URLs, which means the same pic has been uploaded which will not provide additional information and we may want to delete the duplicate pics.

### tweet_json_data

11. We only want original ratings (no retweets), hence 179 retweets can be deleted from twee_json dataframe

## Data Tidiness

1. In the archive dataframe, the *dog "stage"* information (i.e. doggo, floofer, pupper, and puppo) is scattered in four columns and can be merged under one column
2. In the images file, there are three columns with *prediction and prediction confidence information*. Most of the time p1>p2>p3, hence can be combined under one column
3. Merge the three dataframes and bring only required fields in the final dataframe

## Storing

The analysis has been done on the jupyter notebook, wrangle_act.ipynb and the file has been the final merged version of the combined file stored as twitter_archive_master.csv