# wrangling efforts in tweeter data

- **Gathering data**
  - Downloaded from resources
    - twitter-archive-enhanced.csv
    - tweet-json.txt
  - Downloaded programmatically using the Requests library
    - image-predictions.tsv
- **Assessing data**
  - **Quality**

  twitter-archive-enhanced.csv

    - Missing value in expanded_urls
    - Duplicated value in expanded_urls when we have more then photo in tweet
    - tweet_id shold be string not float becuse we didn't make any calc on it
    - timestamp and retweeted_status_timestamp is datetime not object
    - rating_numerator have decimal value in text ignore in column
    - rating_denominator some time not equal 10
    - Malti state for one dog (doggo and puppo) 1 record (doggo and floofer) 1 record (doggo and pupper) 12 record
  - **Tidiness**

  twitter-archive-enhanced.csv

    - Erroneous datatypes for this (doggo, floofer, pupper and puppo)

  image-predictions.tsv.csv

    - img_num belong to twitter-archive not image prediction and related to expanded_urls to display all photo in tweet
    - Erroneous datatypes
      - p1, p2 and p3
      - p1_conf, p2_conf and p3_conf
      - p1_dog, p2_dog and p3_dog

  tweet-json.txt

    - favorite_count and retweet_count belong to twitter-archive
- **Cleaning data**
  - Remove any retweet or reply and drop unusing col
    - 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'

- Missing value in expanded_urls
- Duplicated value in expanded_urls when we have more then photo in tweet
    - complate the missing value in expanded_urls and remove duplicate by recalculate the URL using concat https://twitter.com/dog_rates/status/ + tweet_id
- tweet_id should be string not float
    - convert to string
- timestamp should be datetime not object
    - convert to datetime
- rating_denominator some time not equal 10
    - drop tweet without rating count 15 record
    - Extract from text again
    - Convert to decimal
- rating_numerator have decimal value in text ignore in column
    - Extract from text again
    - Convert to decimal
- Malti record state (doggo and puppo) 1 record (doggo and floofer) 1 record (doggo and pupper) 12 record
    - drop all misclassified
- Erroneous datatypes for this (doggo, floofer, pupper and puppo)
    - rearrange this variable to one variable with value using melt
- img_c convert tweet_id to string
- Erroneous datatypes
    - p1, p2 and p3
    - p1_conf, p2_conf and p3_conf
    - p1_dog, p2_dog and p3_dog
        - convert to one column
- Rearrange Tables to be two tables as "Each type of obs. unit forms a table"
    - Merge tweeter table with favorite and retweet count table