**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