**Data Wrangling Process**

* + - Data wrangling, which consists of:
* Gathering data
* Assessing data
* Cleaning data
* Storing, analyzing, and visualizing your wrangled data.
* Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations.

**Gathering Data for this Project**

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle\_act.ipynb:

1. The Waterdogs Twitter archive.   
   I downloaded the file manually from the website, the file name is [**twitter\_archive\_enhanced.csv**](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv).
2. The tweet image predictions  
    is the second file associated with prediction regarding what the breed of the dog according to a neural network. This file (image\_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the [**Requests**](https://pypi.org/project/requests/) library and the following URL: [**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad\_image-predictions/image-predictions.tsv**](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

1.     The third one is tweet\_josn, we Using the tweet IDs in the Waterdogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's [**Tweedy**](http://www.tweepy.org/) library and store each tweet's entire set of JSON data in a file called tweet\_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas Data frame with (at minimum) tweet ID, retweet count, and favorite count.

*Note: do not include your Twitter API keys, secrets, and tokens in your project submission.*

**Assessing Data**

After gathering Data, I assessed the data both visually and programmatically for quality and tidiness.

By using the pandas functions and methods we get:

.info(), .describe(), .head(), .sample(), .unique() and .value\_counts()

**# Eight (8) quality issues**

**in `df\_twitter\_archive file`**

1- in\_reply\_to\_status\_id column and in\_reply\_to\_user\_id column have many celles with NaN, so in\_reply\_to\_status\_id column has 78 non-null only instead of 2356.

2- in\_reply\_to\_user\_id column has 78 non-null only instead of 2356.

3- retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp columns have 181 non-null olny intstead of 2356.

4- missing values in expanded\_urls column which has 2297 non-null instead of 2356.

5- name column

some values saved as a lowercase ['such', 'a', 'quite', 'not', 'one', 'incredibly', 'mad', 'an',none...etc]

some valuse are non titled like ['BeBe','CeCe','DonDon','DayZ','JD'].

6- We are interested in dogs, however text column shows some tweets are not related to dogs.

7- we have missing data in doggo,floofer,pupper, puppo columns.

8- timestamp , retweeted\_status\_timestamp are saved as object datatype (str) instead of date/timestamp.

9- tweet\_id is saved as int and it will better to be (str).

10-source column wrote in html containg" <" "a" ">" tags.

11- incorrect data in rating\_denominator and rating\_numerator 1776, 960 and 666

12- rating\_numerator should be float not int.

**`image\_prediction file`**

1- 'p1', 'p2', 'p3' inconsistent capitalization (sometimes first letter is capital)

**Overall**

We have challenge in the shape of each file since we have a (2356, 2345, 2075) of 'twitter\_archive','tweet\_json' and 'image\_pred' respectivily.

**# Tidiness**

1- Three data frames `twitter\_archive`, `image\_pred`, and `tweet\_json` should be

one (combined table)

**`df\_twitter\_archive` table:**

1- we have two variables text and short urls,create short\_urls column, drop expanded\_urls.

2- one variable in four columns ('doggo', 'floofer', 'pupper', and 'puppo')

3- we need from tweet\_json these columns only 'id', 'favorite\_count' and 'retweet\_count'

4- rating\_numerator and rating\_denominator columns in twitter\_archive dataset should form one column dog\_rating normalized out of 10.

5- tweet id presents in 'twitter\_archive' and 'image\_predictions' and as id in 'tweet\_info' so we will rename it.

# Clean

1. get a copy from each file to protect the source clean

twitter\_archive\_clean = twitter\_archive.copy()

image\_pred\_clean = image\_pred.copy()

tweet\_json\_clean = tweet\_json.copy()

1. Missing Data
2. Tidiness
3. Quality issues.

## resources

1. Google.
2. 2. Stackoverflow.
3. Pandas websites.