Data wrangling processes:

- Gathering the data
- Assessing the data, you gathered
- Cleaning all issues that documented after the assessment (quality and tidiness)
- Storing the cleaned data in csv or excel or sql file
- Analyzing and make some visualization about your wrangled data
- Make a report on:
    1. Wrangling efforts
    2. Analysis and visualizations

Gathering

There are 3 parts of data must have gathered for this project:

- Twitter archived data from @weratedogs twitter account and I download it manually from the website. It is named as twitter_archive_enhanced.csv
- Image predictions data and it was hosted on Udacity's server and it should be downloaded programmatically by using request library in python with this URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Retweets and favorites of each tweet data and it should be gathered by using twitter APIs. I query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Then i read all the data entire it line by line and save it into pandas data frame.

Assessment

Quality issues:

- In twitter archived data: some columns do not matter
- In twitter archived data: The `text` column shows that there are some tweets which is retweeted tweets and mentions tweets.
- In twitter archived data: Dog names some of them are missed and others are incorrect
- In twitter archived data: These columns (`tweet_id`, `timestamp`, `retweeted_status_timestamp`, `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`) have an incorrect data type.
- In twitter archived data: `expanded_urls` column somtimes have more than repeated URLs and somtimes have different URLs (splitted by ',')
- In twitter archived data: These two columns have invalid and incorrect values `rating_numerator` / `rating_denominator`
- In twitter archived data: `source` column has no variety so it does not matter
- In image prediction data: `tweet_id` column is int data type
- In tweets data: retweets and favorites are float

Tidiness issues:

- timestamp column contains year, month, day, and the time
- These columns (doggo, floofer, pupper, puppo) should become in the same column
- tweet_id column is common in the three data sets

Cleaning

First I take a copy from each data set and assigned to its name 'clean' word.

The cleaning steps as following:

- **Define**: on this step I use verbs to explain what I will do with each issue that I found.
- **Code**: on this step I will convert define to python code to solve the issues.
- **Test**: in this finally step I used some python function and methods or lines of code to test is issue was solved or not.