# Wrangle Data

## Gather

The data is collected from three different sources and they are of different format. They are

1. WeRateDogs twitter archive file, which is a simple comma separated file, read using pandas. read_csv()
2. Data retrieved via twitter API, which is a text file in JSON format, is read line by line using json.loads() and the required attributes are extracted and appended and then converted to a dataframe.
3. Image predictions file, which is in tab separated format and is hosted on an external URL. Instead of using requests library, I have used pandas.read_csv() with separator option as '\t' (tab).

## Assess

1. WeRateDogs twitter archive file
    - Looking at the info() results we could see that the total number of rows are 2356 and all the columns that we are interested in (tweet_id, timestamp, rating_numerator, rating_denominator, doggo, floofer, pupper, puppo) are non-nulls
    - Using duplicated() we could see that there are no duplicate tweet_ids
    - By applying value_counts() on the rating_denominator we could see that 10 is most commonly used. So, will convert the other rows to have a common denominator of 10
    - By applying value_counts() we could see that doggo, floofer, pupper and puppo columns has either None or their value.
    - By applying group by, we could see that there is an overlap between doggo, floofer, pupper and puppo columns i.e one row has values for both floofer and doggo, 12 rows have both pupper and doggo, etc. which needs to be fixed.
    - For some rows that don't have rating_denominator as 10, the text suggests that the ratings were wrongly extracted. So, we could manually correct them.

2. Data retrieved via twitter API
    - All columns have no null values and there's no duplicate tweet_id. From the describe() we could see that both the count columns have valid values from 0 to positive integer

3. Image predictions file
    - All columns have no null values and there's no duplicate tweet_id. From the describe() we could see that both the prediction columns have valid values between 0.0 to 1.0

## Quality

- For all tables, convert tweet_id column to string data type
- Extract correct ratings from text column of archive table
- Remove invalid names in archive table
- Replace underscore '_' with space ' ' for predictions columns
- Trim Source column to have only the href attribute
- Delete retweeted rows i.e. rows with non-null values for retweeted_status_id and in_reply_to_status_id
- In the archive table, check the values of the columns doggo, floofer, pupper and puppo to true or false based on its value
- Update the values of columns doggo, floofer, pupper and puppo for rows manually

## Tidiness

- For archive_df, convert and combine the values for the columns doggo, floofer, pupper, puppo into one column stage.
  - If only one of the columns is set then update the stage column else update it as none.
  - For rows that has more than one stage, update it as multiple
- For archive_df, convert and combine rating_numerator and rating_denominator columns to one column rating

## Other changes

- Find the highest confidence among the three predictions and save it in a new column if it is a dog.
- Merge all three tables to one with only the required columns for analysis.

## Clean

- Save a copy of each data frame for cleaning
- Clean each of the issues and test them.

## Store

- Store the merged dataframe to twitter_archive_master.csv file