wrangle_report

The tweet archive of Twitter user @dog rates, better known as WeRateDogs, was the dataset I worked on wrangling (and analysing and displaying). The whole wrangling process has divided into three main steps.

- Gathering
- Assessing
- Cleaninng

Gathering

The datasets gathered from three different data source:

- **#WeRateDogs twitter archive form Udacity platform**.
  Since the data stored in csv format, I read the data pandas "read_csv" method and stored it in twitter_archive variable. It cointains 17 columns with different datatyps and 2356 entries.

- **Image prediction file from Udacity**
  A programmed download has been used to retrive image prediction file using the Requests library. This file includes information from the WeRateDogs Twitter archive, where each picture was processed by a neural network capable of classifying dog breeds*. This table containing picture predictionswith each tweet id, image URL, and the image number that matched to the forecast with the highest confidence (numbered 1 to 4 since tweets can have up to four images).

- **Twitter API**
  I registered for Twitter's API in order to have access to the retweet and like numbers. This is accomplished by utilising the tweet id from the WeRateDogs Twitter archive, querying the Twitter API for each tweet's JSON data using Python's Tweepy package, and storing each tweet's whole JSON data set in a file named tweet json.txt. The code below allowed us to obtain the required data.

Assesing

The evaluation was conducted visually and programmatically. This is where I intended to identify data quality and tidiness concerns.

Quality Issues:

```
1. (twitter_archive) in_reply_to_status_id, in_reply_to_user_id
datatype falls either integer or string.
2. (twitter_archive) source column is html tagged which is not
readble.
3. (twitter_archive) timestamp, retweeted_status_timestamp should be
```

```
converted to datetime format insted of object.
4. (twitter_archive) rating numerator and rating denominator has
invalid values.
5. (twitter_archive) name entry in name column could be invalid (e.g
a, an, mo is not valid names if it is not tranlated from chinise)
6. (twitter_archive) in_reply_to_status_id, in_reply_to_user_id,
retweeted_status_id, retweeted_status_user_id,
retweeted_status_timestamp, expanded_urls has some missing values.
7.(twitter_archive) lots of None values in dog stages.
8. (twitter_img_pred) There are (2356-2075=281) missing values in
image datsets.
9. (twitter_img_pred) Duplicate tweets has been found on several tweet
ids which is actually retweets.
10. (twitter_archive) Some columns, such as timestamp, have the
incorrect data type.
```

Tidiness Isues:

```
1. dog breeds could be stacked into one coulmn rather than in multiple
columns.
2. unnessary columns which could mege into one column for imgae
predictions.
3. Dog rataing is not standararized.
4. Unnecessary column and rows.
```

Cleaning

In this phase I tried to fix the issues mentioned above to have clean and one maerged dataframe. The process started with copying the fiels into datframes. IN thi sprocess, I tried to follow (Define, code, test) structure. Finally writing into one csv file compltes the whole process.