DATA WRANGLING PROJECT

I. Introduction

Data we are going to use is tweets archive from @dog_rates twitter users aka WeRateDogs.
We have 3 different sources.
The first one, **'twitter-archive-enhanced.csv'** is easily accessible by downloading from course resources. The second one is downloadable from this url "**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**".
The last one can be retrieved using twitter API or simply download from udacity course resources.
Our work consists to gather all 3 datasets, clean them and share some insights.

The gathering phase was pretty easy as the course gave us all the keys to do it.
After the gathering process, I checked carefully each dataset to clean them.

II. Assessing Data

Some quality and tidiness issues has been identified.

A. Quality issues

1. Dog names

After checking dog names, we saw some incorrect names: **a**, **None**.
Most of the tweet start by "This is xxx …" while xxx represent the dog's name. But it appears that some other tweets follow another model. Their first phrase contains "… named xxx.". After that, there still tweets which does not contain dog's name.

2. in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id

Types of these columns is not correct. They must be string instead of float.

3. timestamp and retweeted_status_timestamp

These columns have also incorrect types. We turned them into datetime.

4. rating_denominator

One of the tweets has 0 as denominator. By visually checking the text we found out the correct value.

5. source

Source values are all html content which is not correct for analysis.
We extracted the text from the html content by using BeautifulSoup.

6. doggo, floofer, pupper, puppo

All those 4 columns indicates if the dog is the type of column name or not.
For better analysis, we encoded them as Boolean type. We converted 'None' values as False.

7. favorited, retweeted

favorited means the tweets has been liked once while retweeted means it has been retweeted once. In our understanding, if favorite_count is more than 0 so favorited must be true. Same as retweet_count and retweeted. Some tweets do not follow this logic so we have to correct.

8. Retweets must be removed

retweeted_status_id indicated whether the tweet is a retweet or not.
So based on this column, we removed the retweets.

B. Tidiness Issues

1. *doggo, floofer, pupper, puppo*
   These columns must be combined in only one column named dog_category.
   So we wrote a function to check the category.

2. All the three datasets must be combined in one.
   We did it by using pandas.merge .