Data wrangling: WeRateDogs

Data gathering:

Mainly the data was gathered from 3 sources:

1. Twitter archive data: downloaded from udacity platform.
2. Predictions data: downloaded programmatically using the *request* library.
3. Additional twitter data: accessed through the twitter api *tweepy.*

All these files were then read using pandas' read_csv method.

Data assessing and cleaning:

During the data assessment, the issues found are classified as follows:

- 4 missing data issues
- 4 tidiness issues
- 8 Quality issues

They were dealt with in the mentioned order, below, is a list of the discovered issues and the way they were addressed:

Missing data:

Table	**Issue**	**Solution**
Archive	Missing values for columns: doggo, floofer, pupper, puppo.	- extract dogtypes from the text column and replace the dogtype columns with the extracted ones.
	None instead of Nan in columns name, doggo, floofer, pupper and puppo.	Solved with the previous solution
	Missing 2 records of retweet_count and favorite_count (2354 instead of 2356 rows).	- create a masc for the archive tweet_ids that are in api tweet_ids and use it to select tweet_ids from archive

Predictions	Missing images for some tweets	- create a masc for the archive tweet_ids that are in prediction tweet_ids and use it to select tweet_ids from archive

Tidiness issues:

Table	Issue	Solution
Archive	columns ("doggo", "floofer", "pupper", "puppo") represent the same type of information (dog type)	- write a function that merges a list of non-NaN values - apply this function to get a new column (dogtype) - create a new column in the dataset using this new list - drop the 4 columns
	rating_denominator and rating_numerator represent the 2 components for rating a dog	- create a new column 'ratings' which will be calculated as rating_numerator/rating_denominator. - drop rating_denominator and rating_numerator.
Predictions	The last 9 columns prediction table represent the same type of information(prediction, confidence, isDog)	rename the columns so that you can use pandas' wide_to_long method to melt all 9 columns at once
Api data	This table (api table) has complementary data to Archive table.	use pandas 'merge' method to merge the archive table and api table on tweet_id

Quality issues:

Table	Issue	Solution
Archive	There are 181 retweets (which should not be included in the dataset).	- Only select rows with retweeted_status_id set to Nan
	There are columns which I do not plan to use in my analysis: - in_reply_to_status_id, in_reply_to_user_id - expanded_urls - name - retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp (once retweets	- use the drop method of pandas df.

	are dealt with these columns will be unnecessary)	
	Erroneous datatype in columns: tweet_id, timestamp, dogtype.	- use astype to change: - tweet_id to str - timestamp to datetime - dogtype to category
	Source is given as a html anchor tag	- write a function that recognizes the category of the source - apply this function on the source column values to get a new column - drop the old source column
	rating_denominator and rating_numerator have outliers.	- ratings will be considered outliers if rating >= 7.5 - drop rows that match the above criteria
Predictions	Erroneous datatype for column: tweet_id.	Use astype to convert tweet_id to str
	Some images are very unlikely to be dog images.	- we will consider an image not a dog image if isDog is False with confidence (>0.5). - drop the rows that match the above criteria.
Api data	Erroneous datatype for column: tweet_id.	No need to solve it since api data merged with archive table.

After the listed issued were solved, the result was 2 tables:

- Archive table which has 1991 rows and 8 columns
- Prediction table which has 5959 rows and 7 columns