Wrangle Report

1. Background of the project

This project was done using a tweet archive of a Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

In terms of my data wrangling efforts for this project, I gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, and then cleaned it. I used a Jupyter Notebook, and Python (and its libraries).

2. Data gathering

I gathered three (3) pieces of data from the sources mentioned below:

Data gathered	Source	Number of rows
twitter_archive_enhanced.csv	Manually downloaded as provided by WeRateDogs.	2356 rows
image_predictions.tsv	Downloaded programmatically using the Requests library.	2075 images
tweet_json.txt	Additional data queried via the twitter API using the Tweepy library.	2354 rows

3. Accessing Data

The three (3) data files gathered were loaded in their individual pandas data frames whereby they were subsequently visually and programmatically assessed.

- **Visual assessment:** Data was assessed in an external application (e.g. Excel, text editor).
- **Programmatic assessment:** Data was assessed using pandas' functions and/or methods.

4. Quality issues identified and cleaned

archive table:

1. Replace 109 invalid names in column name of dogs (e.g. 'a', 'an', 'very', 'actually', 'O', 'just', 'my', 'all', 'infuriating', 'the', etc.) with 'NaN'.
2. In the name column, replace the value of None with NaN
3. Remove all rows that have values (not blank or non-null) in retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns, as we only want original dog ratings as per the specification.
4. Drop columns the following columns because were instructed to keep the original tweets that have images:
    - in_reply_to_status_id

- in_reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp

5. Extract only the device name in the source column and delete all other information as it does not provide any helpful information.
6. DataType for column timestamp - Should be datetime
7. Column tweet_id - Should be a string/object as there is no reason for them to be in a numeric form. This will assist when we merge/combine tables.
8. Remove 'HTML tags'/'URLs' that are present in the text column
9. In the rating_numerator column there are values that were not extracted properly (from the text column) as they were decimals

prediction table:

10. Column tweet_id - Should be a string/object to assist when we merge/combine tables.
11. Capitalize the first letter of the names in p1, p2, and p3
12. Drop img_num column as it will not be useful

5. Tidiness issues identified and cleaned

1. archive table - Combine the four columns for dog stages (doggo, floofer, pupper, puppo) to form one column.
2. In the archive_clean table, timestamp column should be split into 3 columns, year, month, and day
3. Combine prediction, and twitter_api tables with the archive table to form one table.

5. Conclusion and Storing data

Data wrangling process was concluded after all data issues identified were cleaned. The cleaning process resulted in single table which had all the information we wanted in order to proceed with our analysis and the data was stored as twitter_archive_master.csv.