Data Wrangling Report

1. **Gathering Data**

The dataset contains tweet archive of Twitter user of WeRateDogs. It is a Twitter account that rates people's dog by making some comments. Most of the ratings have denominator of 10. The images in the dataset were ran in an image prediction neural network and a new dataset was generated. A third dataset was assessed through Twitters' API.

Twitter archive file

A CSV file (twitter_archive_enhanced.csv) was downloaded and imported into dataframe.

Tweet image prediction

Request library of pandas was used to download the tweet image prediction as image_predictions.tsv. this was also imported into a dataframe.

Twitter API data

Pandas Tweepy was used to query data via Twitter API as tweet.json.txt. This was also imported into a dataframe.

2. **Assessing Data**

Visual assessment

The three dataframes were visually assessed for related quality and tidiness issue.

Programmatic Assessment

Pandas commands such as: .sample(), .info(), .unique() and .describe() were used to programmatically assessed the data.

Issues detected

The following were the issues observed with the assessed data:

A. Quality issues:

a. Inconsistency in representation of missing values
b. Erroneous dog names with lower case
c. Text column contains untruncated text instead of displayable text
d. Tweet id is represented as id in tweet.json.txt table
e. Missing records in image_predictions.tsv
f. Html tags in source column
g. Retweets are duplicates and are not original tweets
h. Missing values in expanded urls

i. Timestamp clumn is str instead of datetime
j. Rating denominator values other than 10
k. Records with more than one dog stage

B. Tidiness issues

a. Doggo, floofer, pupper and puppo columns are all about kind of dogs.
b. Columns p1, p1_dog and p1_conf should be two columns; breed and confidence
c. Only one table is required, since all the three tables are all about one observational unit.

3. Cleaning data

The three datasets were cleaned using Define, Code & Test processes. The following cleaning process were done:

a. Replace None with pandas nan
b. Replace erroneous names with np.nan
c. Extract the correct rating fron text and remove rating & links
d. Convert id in tweet.json.txt to tweet id
e. Select tweet id that are also available in image prediction
f. Extract only the text between html tags and convert the datatype to category
g. Take tweets with original tweets and drop those with retweets
h. Drop rows with missing values in expanded url
i. Convert timestap to datetime
j. Change values of rating denominator other than 10 to 10
k. Set rows with more than one dog stage to NaN
l. Merge Doggo, floofer, pupper and puppo into one column
m. Create a new column breed and confidence
n. Merge the three data table into one.

4. Storing Data

The new cleaned and merged dataframe was saved in a csv file named twitter_archive_master.csv