WRANGLE REPORT

The needed datasets were gathered, libraries imported and pandas DataFrames were created to hold the datasets. Afterwards pandas functions (`.info()`, `.describe()`, `.shape()`, `.sample()` `etc.`) were used to check through the datasets for quality issues and tidiness. The following issues were identified:

Quality

t_retweet table (tweet-json.txt dataset)

Issue 1: the id column name in the t_retweet Table was not appropriate.
Cleaning Step: the column was renamed to tweet_id using the pandas `.rename()` function.

t_entwach table (twitter-archive-enhanced.csv)

Issue 2: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns have lots of missing data.
Cleaning Step: the listed columns were removed using the pandas `.drop()` function.

Issue 3: errorneous datatypes (tweet_id, timestamp).
Cleaning Step: the datatypes for the identified columns were changed using the pandas `.astype()` function which allows Dataframe datatypes to be changed.

Issue 4: the following columns (source, name, text) should be renamed (source as tweet_source, dog_name, tweet_text).
Cleaning Step: the columns were renamed using the pandas `.rename()` function.

Issue 5: Drop the dog name column because there are lots of names that are inputted as 'None' and 'a'. This will also not be relevant for analysis.
Cleaning Step: the dog_name column was removed using the pandas `.drop()` function.

Issue 6: Add rating column using the numerator and denominator columns, and then drop the rating_numerator and rating_denominator columns.
Cleaning Step: the rating column was added by dividing the rating_numerator with the rating_denominator, afterwards the rating_numerator, rating_denominator columns were removed using the `.drop()` function.

t_imgpred table (image-predictions.tsv)

Issue 7: the dogs predicted names in column p1, p2, p3 do not have consistent format; should have spaces() in between and not underscore (_)
Cleaning Step: the pandas `str.replace()` function was used to identify the underscore(_) and replaced it with empty space( ).

Issue 8: the names in column p1, p2, p3 should be formatted to be in title case.
Cleaning Step: pandas `str.title()` function was used to change strings to 'Title Case'.

Issue 9: p1_conf, p2_conf, p3_conf do not have a consistent float precision.
Cleaning Step: the floating precisions for the identified columns were changed by using the `.round()` function. The columns were rounded to 6 decimal points.

Tidiness

Issue 10: The variables splited into four columns (doggo, floofer, pupper, puppo) should be in a column named dog_stage.

Cleaning Step: The none values were first of all replaced with NaN and empty string using the `.replace()` function. Afterwards the columns were merged and ajusted as appropriate. The previous four columns were then removed using the `.drop()` function.

Issue 11: tweet_id column data in the tretweet_clean, timgpred_clean and tentwach_clean TABLES will be adjusted to have the same corresponding figures.

Cleaning Step: the columns were divided by the desired number of precison and coverted to integer using `.astype()` function.

Issue 12: Merge the tretweet_clean and timgpred_clean TABLES with the tentwach_clean to create a twitter_achieve_master.

Cleaning Step: the columns were marged using the `pd.merge()` function. It was merged in such a way that only the corresponding column (tweet_id) values were retained.

*****Iteration**

After joining to create the master dataset, rows with null values were dropped using `.drop()` function, and columns with incorrect datatypes were adjusted using `.astype()` function.