DATA WRANGLING REPORT

Gathering the data

- Udacity provided the csv file ‘twitter_master_enhanced.csv’ which was loaded into a dataframe called ‘archive’.
- Using the Requests library, ‘image_predictions.tsv’ was downloaded and read into a dataframe called ‘df_image’
- The json text file provided was used in which the retweets counts and favorite count were gathered and called into a dataframe called ‘df_json’

Assessing the data

Quality issues

- The in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns have empty values.
- Drop the unnecessary columns.
- The datatype of the 'tweet_id' is in integer, and should be converted to string (for all the 3 dataset)
- The source text is embedded in the source url.
- The timestamp is in object datatype, and should be converted to datetime.
- The rating numerator and denominator column have some wrong values.
- Some of the names 'none', 'a', 'an', 'this' etc. are not dog names.
- The dog breeds columns (p1, p2 and p3) have inconsistent case sensitivies.

Tidiness issues

- The dogger, floofer, pupper and puppo columns should all be under one column as it represents only one variable (dog stage).
- Merge the archive data, df_image and df_json into one master data dataset.

Cleaning the data

- Made copies of the original dataset.
- Removed rows with 'retweeted_status_id'.
- Dropped the unnecessary columns.
- From the archive data, the datatype of the tweet_id was converted to string.
- The text is embedded in the source url was extracted.
- The timestamp was converted to datetime.
- The ratings were compared with the texts and wrong values were replaced. Also, ratings with decimals that were incorrectly extracted were fixed.
- Some of the unsual dog names were dropped and ‘None’ changed to unknown.
- The dog breed columns were standardized by changing the values to lowercase.

Storing the data

- The clean data was saved into a csv file ‘twitter_archive_master.csv’.