Wrangle Report

During this project I have been working with data gathered from Twitter account @dog_rates. The process included gathering additional data, assessing them visually and programmatically, cleaning and storing cleaned data in twitter_archive_master.csv, and analysis.

Data Gathering

Gathered data included three files:

- Provided by Udacity csv file 'twitter_archive_enhanced.csv' for manual downloading. It contains basic information about 2356 tweets (like tweet_id, timestamp, tweets text).
- File 'image_predictions.tsv' was available for programmatic download. It contains 2075 predictions made by a neural network that can classify breeds of dogs.
- The third piece of information had been gathered using Twitter API and Python's Tweepy library. File 'tweet_json.txt' contains information about the amount of likes and retweets each tweet received, and has info about 2327 tweets.

Data Assessing

I used two types of data assessment:

- Visual assessment: using .head(), .tail() and .sample() methods I displayed data from each file in the Jupyter Notebook.
- Programmatic assessment: .info(), .value_counts() and .duplicated() are examples of pandas' methods used to assess the data.

Tidiness issues:

1. Breeds and confidence levels for the breed predictions column from the image prediction table should be part of the tweets table.
2. Engagement table (retweet_count and favorite_count columns) should be part of the tweets table.
3. 4 columns (doggo, floofer, pupper, puppo) are categories of dog 'stage' and need to be one column 'stage' with 4 categories: doggo, floofer, pupper and puppo in it.

Quality issues

1. Number of observations differ in twitter-archive-enhanced(2356), imgage_predictions(2075) and tweet_json(2327).
2. Set up tweet_id column as index in tweets_master table for work confinience.
3. Some rows contain more than one dog stage
4. Missing values in doggo, floofer, pupper and puppo columns are strings 'None' instead of NaN.
5. Redundant retweet rows.
6. Redundant columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls' in tweets table.
7. Wrong data type of 'timestamp' (object instead of timestamp) column.
8. Inconsistencies in the 'rating_denominator' column need to be investigated and fixed if possible. Issues might have happened during text parsing and though I'll display text to check it.
9. There are inconsistencies in the 'rating_numerator' column (numerator is greater than denominator, too great values). 'Numerator is greater than denominator' is mainly a feature of WeRateDogs account, but partly might be an issue during text parsing.
10. Upper and lowercase breed names.

Data Cleaning

As a first step I made copies of original pieces of data:

- tweets_clean = tweets.copy()
- image_pred_clean = image_pred.copy()
- engagement_clean = engagement.copy()

Next I decided to fix tidiness issues first so:

- Merged breeds and confidence levels for the breed predictions column from the image prediction table, engagement table and the main tweets table. As a base column I used tweet_id column which I later transformed to index.
- To make column ‘stage’ with 4 categories: doggo, floofer, pupper and puppo in it from 4 columns (doggo, floofer, pupper, puppo) I cleaned 2 quality issues:
    - Dealt with rows containing more than one dog stage by displaying tweets text
    - Changed string input ‘None’ to NaN in rows having missing values

The result of the cleaning efforts above is DataFrame tweets_master. Further cleaning process was performed on this joined DataFrame:

- Dropped redundant retweet rows.
- Dropped redundant columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'.
- Changed data type of ‘timestamp’ column from object to timestamp.
- Cleaned up inconsistencies in the 'rating_denominator' column by displaying tweets text. Issues might have happened during text parsing but partly it is a feature of WeRateDogs account.
- Cleaned up inconsistencies in the 'rating_numerator' column (numerator is greater than denominator, too great values).
- Change all breeds to lowercase.

After I cleaned all the issues above I stored data in ‘twitter_archive_master.csv’. The file contains 2148 observations.