Project – Data Wrangling (We Rate Dogs Twitter Data)

The "We Rate Dogs" Twitter archive contains over 5000 tweets, which have been filtered to create the enhanced archive that forms the basis of this analysis. The goal of this project is to wrangle the data - gather, assess, and clean - into a tidy dataset, and then provide analyses and visualizations.

Gathering Data:

1. **Twitter Archive Data:**
   Already made available by Udacity by the name, ‘twitter-archive-enhanced.csv’

2. **Image Predictions Data:**
   Downloaded programmatically using the link 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

3. **Twitter API Data:**
   Queried the Twitter API to get each tweet's json data and stored that JSON Data into a pandas data frame

Assesing Data:

1. **Twitter Archive Data:**
   The following was assessed:
   Quality Issues:
   - 181 retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
   - 78 replies (in_reply_to_status_id, in_reply_to_user_id)
   - Timestamp field is in string format (object)
   - Some of the rows have invalid strings in the name column, e.g. "a", "an", "in". These words are all the 3rd word in the tweet.
   - Values of "None" in the name column.
   - There are 23 cases where the denominator of rating != 10. These entries will be removed.
   - Rename columns with more appropriate names: "timestamp" to "tweet_timestamp", "text" to "tweet_text", "rating_numerator" to "dog_rating_out_of_ten", "name" to "dog_name"
   - Since retweets and replies will be removed, the column "retweeted_status_timestamp" will be removed as it will no longer provide any useful information.
   - Remove column "rating_denominator" once all the values that != 10 have been removed since this will no longer provide any useful information.

Tidiness Issues:

- Columns related to retweets are not applicable for original tweets
- Columns related to replies are not applicable for original tweets
- The columns with numerical data that are typically used for analysis are located to the far right of the table, and the columns with long strings are on the left; this makes it difficult to readily see the data that will be used for analyses.
- Change columns "doggo", "floofer", "pupper", and "puppo" from wide to long format.

2. **Image Predictions Data:**
   The following was assessed:
   Quality Issues:

   - Entries where the first (i.e. most confident prediction) has a False value for "p1_dog" can be removed.
   - The "p1" and "p1_conf" columns will be renamed with more explanatory titles.
   - The column "jpg_url" will be removed since url data is already contained in the twitter archive data
   - The "p2" and "p3" related columns will be removed as using the most likely prediction ("p1") in analysis
   - After removal of "False" entries, the "p1_dog" column will be removed as it will no longer add any valuable information.

3. **Twitter API Data:**
   The following was assessed:
   - After trying to merge the data, it appears that there is some non-numeric values for the "tweet_id" inputs which will need to be removed.

Cleaning Data:

1. **Twitter Archive Data:**
   Make a copy of the data and perform operations on that.
   - There are 181 retweets and we are interested in original tweets
     Define: Drop all rows containing retweets, where these columns will be non-null: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.
   - There are 78 reply tweets; we're only interested in "original tweets".
     Define: Drop all rows that are replies, those that have non-null values in these columns: in_reply_to_status_id and in_reply_to_user_id.

- All columns related to “retweets” will be empty (we're not interested in retweets).
  Define: Since we don't want retweets, we can drop all columns related to retweets: retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- All columns related to “replies” will be empty (we're not interested in replies).
  Define: Drop all columns related to replies: in_reply_to_status_id and in_reply_to_user_id.
- The timestamp column is in string format, it's the wrong data type.
  Define: Convert timestamp to datetime data type.
- Define: Removing multiple cases of where the denominator of rating != 10.
- Define: Rename columns with more appropriate names: "timestamp" to "tweet_timestamp", "text" to "tweet_text", "rating_numerator" to "dog_rating_out_of_ten", "name" to "dog_name"
- Define: Removing column rating_denominator
- Define: Change columns "doggo", "floofer", "pupper", and "puppo" from wide to long format.

- There are many tweets with regular words in the name column that are NOT a valid name.

  Define: Replace all lowercase words in the name column with the string "none".

- Reorder the column placement: bring numerical columns to the left.

  The coding and Testing part of all these have been given in the Jupyter Notebook, ‘wrangle_act.ipynb’.

2. **Image Predictions Data:**
   Make a copy of the data and perform operations on that.
   - Entries where the first (i.e. most confident prediction) has a False value for "p1_dog" can be removed.
   - The "p1" and "p1_conf" columns will be renamed with more explanatory titles.
   - The column "jpg_url" will be removed since url data is already contained in the twitter archive data
   - The "p2" and "p3" related columns will be removed as using the most likely prediction ("p1") in analysis

- After removal of "False" entries, the "p1_dog" column will be removed as it will no longer add any valuable information.

The coding and Testing part of all these have been given in the Jupyter Notebook, 'wrangle_act.ipynb'.

3. **Twitter API Data:**
   Make a copy of the data and perform operations on that.
   - Finding non-numeric values for "tweet_id"

     The coding and Testing part of all these have been given in the Jupyter Notebook, 'wrangle_act.ipynb'.

Combining the Data:

Combining the three dataframes into one before analysing has also been performed. Ultimately the combined data and all the separately cleaned data has been saved in csv files.