Data Wrangling Explanations

The following issues were identified during the data wrangling process and was cleaned.

Quality issues

- Tweet Archive dataframe *
- Unoriginal tweets are included
- Redundant columns in the archieve dataframe
- Decimals ratings and tweet more than 1 ratings mentioned were truncated wrongly
- The source column is not properly formatted
- Incorrect datatype for tweet, timestamp colums
- Inconsistent dog names
- Incorrect entries at the expanded URL column

Tweet Json dataframe

- Retweet data
- Incorrect datatype for tweet

Image definitions dataframe

- Incorrect data type for tweet #### Tidiness issues
- The dog categorizations can be merged into a single column
- Merge the 3 dataframes for proper analysis

Summary of the Solutions to each of the issues

- Unoriginal tweets are included : Rows with data in the retweeted_status_user_id column was dropped from the dataframe
- Redundant columns in the archieve dataframe: The following columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_timestamp)were dropped using the drop function.
- Decimals ratings and tweet more than 1 ratings mentioned were truncated wrongly: The text column was iterated for each row and the ratings pulled using the regex library. Tweets with more than one ratings mention, the later rating was adopted. The regex profile was also able to capture ratings that are decimal and replaced the values in the rating_numerator and rating_denominator columns with the fetched correct values
- The source column is not properly formatted: Regex was used to filter the source information for each row and replaced with the wellformatted result
- Incorrect datatype for tweet, timestamp columns The tweet_id column on all dataframes were converted to string and the timestamp column converted to datetime datatype
- Inconsistent Dog names: All dog names that started in small letters were made None
- Incorrect entries at the expanded_url column: The tweet_id was merged with a consistent url to fill up the expanded_url column
- Retweet data in tweet_json dataframe: Columns that have True for retweet were dropped

- The dog categorizations can be merged into a single column: A new column named dog_stage was created and initialized with None, each row of the column was iterated and based on the values of the doggo, floofer, pupper, puppo columns, a new value is set
- Merge the three dataframes All the 3 dataframes were merged on the tweet_id column and rows with null values were dropped

In [ ]: