Wrangle report of weRatedogs tweets

Datasets Description

There are three sources of data in this project.

1. Basic archived data from weRateDogs, aka archive_df.
2. Dog breed prediction was provided by a volunteer who used a neural network classification process on images attached to tweets, aka image_predictions.
3. Live data tweets are downloaded using the Twitter API, aka tweet_df.

Issues in each dataset

1. The tweet_df had one quality issue; the tweet_id should not be an int datatype. However, the data frame was well structured since only the tweet_id, retweet_count, and favorite_count was extracted.

2. The image_predictions data frame had one quality issue; the tweet_id should not be an int datatype. The data frame was not well structured as there are multiple rows of predictions, prediction confidence, and the likelihood of prediction being True and False. Each of these values needed to be merged to make better sense of the data.

3. The archive_df data frame had several quality and tidiness issues.
   The tweet_id should not be an int datatype.
   The entries for retweeted tweets should be dropped as this analysis focuses on original tweets.
   The replies to tweets should also be dropped.
   The timestamp column should be a Datetime datatype, not an object.

The rating_denominator must be set to 10 because, as stated earlier, the denominator is usually 10.

The data frame was poorly structured, with multiple rows of dog stages. The Doggo, Floofer, Pupper, and Puppo columns need to be merged to make better sense of the data.

Quality Issues

1. The timestamp column of the archive_df was changed from an object datatype to a Datetime data type.
2. I changed the source of tweets column in archive_df into a more readable format by replacing the links with short texts, including iPhone, Vine, Twitter web, and Tweetdeck.
3. I replaced the irrelevant names with None.
4. I dropped the rows of “None” dogs' names.
5. Since the focus was on original tweets, I set all the retweeted tweet IDs to null.
6. Since the focus was on original tweets, I set all the reply to tweet IDs to null.
7. I changed the data type of tweet IDs in the combined data frame from int to object data type.
8. I dropped tweets that did not have the value of rating denominator equal to 10.

Tidiness Issues

1. In the image_predictions table, the multiple rows of predictions, prediction confidence, and the likelihood of the prediction being True and False were merged into two columns, “breed” and “confidence” This made the information in the columns more understandable.

2. In the archive_df data frame, the Doggo, Floofer, Pupper, and Puppo columns were merged to make better sense of the data. Each of these was merged into a single column, “dog_stage.” This made the table more presentable.
3. The 3 data frames: tweet_df, image_predictions, and archive_df, were combined to form a more consolidated data frame. This was done to give a better presentation of the data for further assessment and analysis.
4. In the combined data frame, all the associated columns with retweets and reply to tweets, including: 'in_reply_to_status_id,' 'in_reply_to_user_id,' 'retweeted_status_id,' 'retweeted_status_user_id,' and ‘retweeted_status_timestamp' were dropped.