Wrangling

In this project, the data were gathered from three different sources. The first CSV file contains the bulk data from tweets of the account WeRateDogs (@dog_rates). The texts were pre-processed and saved in the 'twitter-archive-enhanced.csv' CSV file. A second file contains the results of image analysis saved in a TSV file named 'image-predictions.tsv'. the last source is the current (2021-Feb-04) status from the describer tweets, which were gathered using the ‘tweepy’ API. From the status was used ‘favorite_count’ and ‘retweet_count’ information. These three data were stored in a three separated Pandas DataFrames, ‘twitter_archive_df’, ‘image_predictions_df’, and ‘tweet_statuses_df’, respectively.
In the assessment part of this project the following issues were found:

Regarding Quality

Enhanced Twitter Archive table

- ‘timestamp’ and ‘retweeted_status_timestamp’ columns type are string, not datetime type.
- ‘in_reply_to_status_id’, ‘in_reply_to_user_id’, ‘retweeted_status_id’, and ‘retweeted_status_user_id’ are are of type float and should be of type string.
- Nulls represented by the string 'None' in 'name', ‘doggo’, ‘floofer’, ‘pupper’, and ‘puppo’ columns.
- Strange rating_denominator values (ie. 0, 2, 7, 11, 15...). Same for rating_numerator.
- Unoriginal ratings (retweets).
- Unnecessary columns

Image Predictions table

- Missing information: ‘image_predictions_df’ has 2075 rows and ‘twitter_archive_df’ has 2356 rows ***(can't clean)***
- Same tweets are not dogs (according to the prediction).
- Unnecessary columns

Tweet Stuatus Counts table

- Missing information: ‘tweet_statuses_df’ has 2308 rows and ‘twitter_archive_df’ has 2356 rows ***(can't clean)***
- Inconsistent ‘tweet_id’ column datatype

Regarding Tidiness

- ‘pupper’, ‘puppo’, ‘doggo’, and ‘floofer’ should be a single column of type category
- All tables forms a single observational unit. Therefore, they can be brought together as single data frame.

To clean the data, the three DataFrames were copied, so that the originals are kept as a reference.
The unnecessaries columns were deleted using the Pandas' 'drop' method. Pupper, puppo, and doggo columns were collapsed int a new 'age' column. The columns with the wrong datatype were change using Pandas' 'astype' method and 'to_datetime' function. The DataFrame were merged using Pandas' 'merge' function. And the wrong cell's content (Nullrepresenteded by 'None') were corrected using Pandas' 'replaced' method.