We Rate Dogs

The We Rate Dogs Twitter feed provided ample opportunity to practice some data wrangling and cleaning skills; the steps of which are outlined here.

Begin with the missing data. Gather using the Twitter API by creating a query to save the JSON into a text file, and make a dataframe consisting of the tweet ids, and the retweet and favorite counts.

The image prediction file and the enhanced archive file are provided.

Upon visual and programmatic assessments, several quality issues and a few tidiness issues can be identified:

Provided archive table, labeled df_1
- tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_id columns are stored as int but not needed for any mathematical operation
    Convert datatype for described columns to string with `.astype(str)`.
- timestamp is not datetime
    Convert datatype for described columns to datetime with `pd.to_datetime`.
- tweet 313, 835246439529840640, has a denominator of zero and two others have low denominators, possibly erroneous
- some tweets have significantly larger numerators, possibly erroneous
- within tweet text, numbers with decimals found; dtype for rating_numerator and rating_denominator should be float
    Re-do the rank extractions from tweet text using regular expressions and `str.extract`, and store as float datatype. Filter out absurd ratings, using box plots.
- rows with values in retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp are not original tweets
    Use `isnull` to filter out unoriginal tweets, then drop the retweet columns.
- incorrect/absurd names values
    Set name column values as `nan` wherever there is no valid name in title case. Drop rows without valid names.
- some dogs are simultaneously in multiple stages, but it's not reflected in the columns
    Search the tweet text again for stages, store in the columns, then collapse into one column.

Wrangled table, labled tweet_data
- tweet_id stored as int, should be string
    Convert datatype for described columns to string with `.astype(str)`.

Provided image prediction table, labeled image_prediction

- tweet_id stored as int, should be string
    Convert datatype for described columns to string with `.astype(str)`.
- not all predictions are dogs
    Eliminate rows where there is no valid dog prediction and keep only the most likely dog breed prediction.
- inconsistent capitalization in p1, p2, p3
    Apply `.str.lower()` to predicted_breed column.

Tidiness
- Floofer/Doggo/Puppo/Pupper columns: column headers are values, not variable names
    This was addressed earlier upon noticing some dogs were in multiple stages.
- expanded_urls sometimes have more than one url
    Use `str.split` to separate urls
- a single observational unit (tweet) is stored in multiple tables
    Join the three tables by `tweet_id`. Use "inner" to join the cleaned tweet data to the cleaned provided archive file, but then join this new dataframe with the cleaned image prediction file on a left join, so as not to lose quality tweets just for missing images.

There is now one quality, tidy dataframe for the We Rate Dogs tweets, ready for analysis.