WeRateDogs Twitter Dataset wrangling and analyzing project

The Dataset was obtained from three sources:

- Enhanced Data Archive
- Image predictions file from programmatic download using the requests library.
- Twitter API for additional data (obtained from the notebook and read into a data frame

Visual and programmatic assessments were done to identify quality and tidiness issues.

Then, copies of the data frames were made before cleaning began.

The Enhanced Data Archive was cleaned through the following:

- Removing inconsistent values in the name column
- Replacing missing values in the name column recorded with np.nan
- Dropping the retweeted_status_id column
- Dropping the retweeted_status_user_id column
- Dropping the retweeted_status_timestamp column
- Correct datatype for timestamp from object to DateTime
- Dropping the in reply_to_status_id column
- Dropping the in reply_to_user_id column
- Extracting rating scores correctly from tweet text using RegEx and converting it to float
- Extracting tweet source from source and converting it to category datatype
- Created dog_stages column and removed (doggo, floofer, pupper, puppo) columns.
- Dropping rows with values other than 10 for rating_denominator to have uniform data
- Removing rows with invalid names in the dataframe
- Removing ratings and links from the text column using RegEx.

The Image Predictions table was cleaned through the following:

- Updating inconsistent values format - capital and small letters in p1 column
- Updating inconsistent values format - capital and small letters in p2 column
- Updating inconsistent values format - capital and small letters in p3 column
- Condensed the p1, p2, p3, and p1_conf, p2_conf, and p3_conf to get the breed and confidence columns with the highest confidence predictions and drop other columns
- Dropping Image number columns in (the predictions table)

The Twitter API table was cleaned through the following:

- id column name, not consistent with the other two tables, was changed to tweet_id
- Rows with retweets were removed
- Specific columns were selected - tweet_id, favorite_count, retweet_count

After cleaning, a master dataset was obtained by combining the three datasets together using the merge function in pandas.

Thank you.