Wrangle report

Firstly, we gathered the data and loaded in Jupyter Notebooks. Afterwards, we proceed to assess the gathered data for quality and tidiness issues.

For the three files are read, it is used:

- info() to have a concise summary of the DataFrame;
- describe to generate descriptive statistics of DataFrame columns;
- analysis of duplicates, missing values and unique values;
- the first and last 10 rows using head() and tail();

We could note that:

The info() method reveals several quality and tidiness issues:

- There are 181 retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- There are 78 replies (in_reply_to_status_id, in_reply_to_user_id)
- There are 2297 tweets with expanded_urls (links to images) hence, 59 tweets with missing data
- The timestamp field had string format (object)
- There are 4 columns for dog stages (doggo, floofer, pupper, puppo) and wrong category for stages
- The columns related to retweets are not applicable for original tweets
- The columns related to replies are not applicable for original tweets

The head() and tail() methods show us several issues:

Quality:

- The timestamp column has dates in string form.
- Some of the rows from the tail() output above have invalid strings in the name column, e.g. "a", "an", "in".
- Values of "None" in the name column.

Tidiness:

- The columns with numerical data that are typically used for analysis are located to the far right of the table, and the columns with long strings are on the left; this makes it difficult to readily see the data that will be used for analyses.

Secondly, we assess the columns in depth, using the same techniques. The findings include:

- Assuming that the dog names are all capitalized, words beginning in lowercase are not names.
- For the 59 tweets that contain missing data in expanded_urls, 56 are replies or retweets. The remaining 3 tweets (at indexes 375, 707 and 1445) with NaN in the expanded_urls column all have valid ratings but no urls within the text column.
- For rating_numerator and rating_denominator, the describe() method shows us some quality issues:
    - The max values are huge: 1776, 170. The minimum is 0 for both.
- There are 4 types of sources, and they can be simplified by using the display string portion just before the final "<\a>":
    - Twitter for iPhone
    - Vine - Make a Scene
    - Twitter Web Client
    - TweetDeck
- Over 500 instances where the algorithm did not predict a dog breed from the image; Combined, there are 324 cases where there is no valid dog breed from any of the three predictions.

Hence, we proceed to clean the data following the schema below:

Quality

- Archive
    1. Drop rows with non-null values in 'retweet' data
    2. Drop rows with non-null values in 'replies' data
    3. Drop nulls for 'expanded URLs'
    4. Reformatting 'timestamp' to DateTime
    5. Invalid values in 'name' to "None"
    6. Manually fix of ratings and dropping denominators =! 10
    7. Drop numerators > 15
    8. Establish source category
- Images Prediction
    1. Solve missing values when merging data ### Tidiness
- Archive
    1. Create one column for dog stage and drop invalid category ('floofer')
    2. Drop 'retweet' columns
    3. Drop 'replies' columns

    4. Drop 'rating_denominator' and rename 'rating_numerator' as 'rating'
- Json-data
    1. Rename columns to allow merge smoothly
- Merge json-data and archive
- Images Prediction
    1. Create 'breed' and 'confidence'; merge to archive
    2. Capital letter to 'breed'
- Order columns