Data wrangling process for WeRateDogs tweet data by M. Joukamaa

The task was to gather, assess, and clean data pertaining to the tweet archive of the @WeRateDogs Twitter account. This data came from three sources: 1) a .csv file containing the archived tweets with most of the tweet content and attributes as well as the names, ratings, and stages of the dogs featured in the tweets: 2) a .tsv file containing neural net predictions of the dogs' breeds based on the archived tweet images; 3) retweet and like counts for the archived tweets, gathered directly from Twitter.

The two flat files containing project data were downloaded from a cloud server using Requests library and then saved locally and read into pandas dataframes. Data from Twitter was extracted using an API object created with the tweepy library and written into a text file. The tweet data in the text file was then read programmatically and converted into a dataframe.

The data in these three dataframes was then assessed both programmatically and visually using `.info()`, `.head()`, `.tail()`, `.sample()`, and `.value_counts()` pandas functions. As suggested in the Project motivation concept, there were several issues in the data that required cleaning. Following quality issues were found that were relevant to further analyses to be made with the data:

- Tweet id in integer datatype in tweet and prediction dataframes
- Timestamp not in datetime format
- Dog stage data not in categorical datatype
- Name column including words that are not names (eg. 'a', 'an', 'the', 'very')
- Dog name and stage columns denoting missing data with 'None' values instead of 'NaN'
- Rating numerator including several inappropriately high (20-1776) values
- Rating denominator including several values other than 10
- Tweet dataframe including a large number of retweets

- Tweet dataframe including a large number of tweets that have no images

Also, two major tidiness issues with the data were identified:

- Dog stages (doggo, puppo, pupper, floofer) classified into separate variables
- Information about two different observational units - ie. the tweets and the dogs themselves - scattered across three different dataframes

The quality issues were cleaned using methods such as datatype conversions, filtering by null and non-null values, recalculating anomalous rating numerator and denominator values, and reassigning the ‘None’ values to ‘NaN’ values with the `np.nan` function. The dog stage tidiness issues was cleaned by using pandas’ `.combine()` function, while the proper division of observational units was achieved by appropriate merging and dropping of columns.

Finally, the two cleaned dataframes - one containing the tweet data and the other one containing the dog data - were saved into .csv files using `.to_csv()`.