WeRateDogs Data Wrangling Outline

The data wrangling efforts in this project include gathering, assessing and cleaning data. The following steps were performed for each part.

Gathering Data

1. One of the files ('twitter-archive-enhanced.csv') in this project was provided as a csv file, which was read into a pandas DataFrame.
2. Image prediction data was imported via ‘requests’ library using the provided link and read into a pandas DataFrame.
3. Additional tweet data was accessed through twitter API via ‘tweepy’ library. A twitter developers account was required to access the twitter API. Json object for each tweet id was appended to a .txt file. The required tweet data was accessed through this .txt file and added to a new pandas DataFrame.

Assessing and Cleaning

1. The three previously gathered pandas DataFrames were merged into one master pandas DataFrame.
2. Rating numerators were re-extracted from tweet texts using regex, converted from strings to floats, and any rows with rating numerators greater than 15 were dropped. Some rating numerators were corrected manually, since the regex picked up the wrong numbers from the tweet texts. Some rows were dropped, since their tweets did not contain any ratings at all.
3. Rating denominators were not all equal to 10, so they were all set to 10.
4. A lot of lowercase strings were present in the ‘names’ column, which were not actual names. Those entries were converted to nulls.
5. The `name`,`doggo`, `floofer`, `pupper`, `puppo ` columns contained the string “None” instead of formal null values. These strings were converted to np.nan nulls.
6. Dog stages were represented in 4 different columns contrary to data tidiness requirements. The 4 columns were concatenated into one column and a comma was inserted between any concatenated strings. All empty values were replaced with nulls and the original 4 columns were dropped from the DataFrame.

7. Many rows contained non-null retweeted status id data, which indicated that these were retweets. These rows were dropped.
8. Any unneeded columns were dropped from the DataFrame.
9. Any non-descriptive columns were renamed.
10. Any rows with missing image prediction or favorites data were dropped from the DataFrame.
11. The cleaned DataFrame was saved to a new csv file.