WeRateDogs Twitter Archive - Wrangle Report

In this report I outline the wrangling efforts to wrangle the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering:

Data was gathered from 3 sources:
1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The `favorite_count, retweet_count and followers_count` were extracted programmatically from this file.

Assessment & Cleaning:

After visual inspection and checking the data of the three files programmatically using Jupyter Notebook and pandas' functions, I identified several issues of quality and tidiness as follows:
Quality:

- The name column represents the dog names which are all capitalized, so words that begin in lowercase are not a dogs' name as ("a", "the" and "an").
- There are 181 retweets (retweeted_status_id) and there are 78 replies (in_reply_to_status_id).
- There are only 2297 tweets with linked to images; so, 59 tweets are without images.
- The following redundant columns (in_reply_to_user_id, 'retweeted_status_user_id', 'retweeted_status_timestamp') may needs to be droped
- The data type for tumestamp is string and it contains '+0000' string.
- The rating_numerator column has incorrect values as (70, 7, 150, 11, 2).
- The rating_denominator column has incorrect values as (0, 2, 170, 6).
- The expanded_urls column has some repeated URLs in the same cell.
- There are 2075 image predictions only, the rest will be classified as "missing data".

Tidiness:

- There are 4 columns for dog breed (doggo, floofer, pupper, puppo) instead of one "dog_stage"
- The columns' names are not descriptive.
- The json_data table should be combined with the archive table.
- There are three columns to indicate is_dog instead of one in the image_prediction table.

Data Cleaning:

- For each quality and tidiness issue, the work flow was
    1- Define the cleaning steps.
    2- Write the appropriate code to clean.
    3- Test the results.
- After cleaning the clean data was saved in a comma separated values (.CSV) file.