WeRateDogs Twitter Archive - Act Report

WeRateDogs Data Gathering:

For this project, data was gathered from 3 distinct sources:

- WeRateDogs Twitted enhanced archive was manually downloaded from Udacity course material
- The image predictions file which was programmatically downloaded from Udacity servers
- I attempted to download the json data for tweets published by the WeRateDogs handle through the **tweepy** API. However, since I was not able to successfully create a Twitter developer account. I used the **twee-json.txt** file provided in the course material.

The three source data files were loaded into separate dataframes: archive, predictions and json_data

WeRateDogs Data Analysis and Cleaning:

I began the data assessment by viewing the information on the **archive** table to identify data quality and cleaning issues:

- There were 181 retweets (**retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp** NOT null) and 78 reply tweets (**in_reply_to_status_id, in_reply_to_user_id** NOT null) which may lead to counting the same dog multiple times. All these records were dropped as part of the cleaning exercise

- The timestamp column was in string format and it was converted to datetime

- There were 109 tweets with regular words in the name column that are NOT a valid name; these words are always the 3rd word in the tweet and are all lowercase; all valid names start with an uppercase letter. Hence, all lowercase words in the name column were replaced with the string "none".

- Records with missing data in **expanded_urls** field were dropped
- The **rating_numerator** and **rating_denominator** columns were checked for value ranges; 4 records were found where the tweet text contained the correct rating. For these records, the rating was manually corrected
- I also found 13 tweets about multiple dogs/pups and hence **rating_denominator** is NOT equal to 10. I handled these records by normalizing the **rating_numerator** and changing the **rating_denominator** field to 10. This also needed changing the data type of **rating_numerator** to float
- Tweets with exceptionally large **rating_numerator** values were dropped, as the text didn't contain a valid rating (# out of 10)
- The **source** column values were cleaned by extracting the text from the html tags
- The 4 dog stage columns - doggo, floofer, pupper, puppo were combined into the stage column. For tweets which did not mention stages, the field **stage** was set to 'none'
- The **retweet_count** and **favorite_count** columns from the **json_data** table to the archive table, joining on tweet_id
- Dog breed names were not standardized in the predictions table, some is capitalized and other lowercase. We dealt with this by:
    - Converting all names to lower cases
    - Converting all spaces into an underscore
    - Converting all dash to underscore

- The best dog breed prediction and associated confidence level fields were combined with the archive table; these pieces of information provide additional data about the dog in the tweet based on the tweet's image
- The **tweet_id** field was integer data type. However, **tweets_ids** cannot be added or subtracted and no arithmetic operations can be performed on tweet_id values. Hence the data type was changed to string

The cleaned columns in the archive table were reordered so that numerical columns were not to the extreme right, preventing visual assessment. The cleaned table was saved to the new "**twitter_archive_master.csv**" file.