WeRateDogs Twitter Archive - Wrangle Report

- **By Pradeep Gurunathan**

In this report I summarize the wrangling efforts to collect and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering

Data was gathered from three sources and stored in separate files:

1. Manually downloaded WeRateDogs Twitter Enhanced archive file from Udacity server and saved as "archive_data".
2. Programmatically downloaded image predictions file from Udacity server and saved as "predictions_data".
3. Using twitter API and tweepy library tweet's JSON file was downloaded programmatically and saved as "json_data".

These three data files are stored separately.

Assessing and Cleaning

- Both by visual and programmatically data was assessed to identify quality and tidiness issues.
- First "archive_data" was assessed first
    1. All rows containing non-null values in the retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp , and also in the in_reply_to_status_id and in_reply_to_user_id columns were dropped, as per the requirements. These columns were then also dropped.
    2. timestamp column was converted to datetime data type
    3. 4 dog stage columns were melted into the stage column; tweets without stages were set to 'none'. Several had 2 stages set, so I kept only the one with the lower overall count.
    4. rating_numerator and rating_denominator columns were checked for value ranges; I decided to keep only tweets with single ratings. Several tweets' ratings were manually corrected with values from the text. Tweets with large numerators were dropped, as the text didn't contain a valid rating (# out of 10). After the ratings were fixed, I dropped the rating_denominator column (it contained only '10's) and renamed the rating_numerator column to rating .
    5. odd words in the name column were replaced with 'none'.
- Predictions table itself was not cleaned. There were many tweets with no dog breed predicted, these were left as is. The best prediction for breed and associated confidence level were extracted and merged into the archive table.
- json_data table itself was not cleaned. The retweet_count and favorite_count columns were merged into the archive table, and the data type reset to int. One tweet was missing both counts so was dropped.
- remaining cleaned columns in the archive table were reordered, then the table was saved to the new "twitter_archive_master.csv" file. The predictions and json_data tables had not been cleaned, so were not saved.