Data Wrangling Report

This report is made to document all the steps taken in cleaning the **WeRateDogs Twitter Archive**. This includes gathering data from the given sources as well as steps taken to clean and merge relevant sections of the dataset(s).

Data Gathering:

Data was gathered from three sources:

1. **The WeRateDogs Archive**. This dataset was provided by Udacity and was downloaded from its servers.
2. The **Image Prediction** dataset which was downloaded programmatically from the Udacity server as well.
3. The raw data for each tweet present in the JSON format. It was mined using the Twitter Developer API and from this file the important variables: **Tweet ID**, **Retweet Count** and **Favorite Counts** were extracted from this file.

These three datasets were loaded into my Jupyter Notebook workspace and saved as twitter_enhanced, images and status respectively.

Data Assessment and Cleaning:

The Twitter Archive dataset was first assessed for data quality and tidiness issues. First I looked at the different data types of the variables in the data. Several of these issues were noticed and taken care in the following steps:

1. All columns relating to retweets and replies (**retweet_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp**, **in_reply_to_status_id**, and **in_reply_to_user_id**) were dropped as they played no part in our analysis.
2. The timestamp column data type was changed from ‘object’ to ‘Date’.
3. Source of tweets were extracted from the HTML strings in the **source** column.
4. The **rating_numerator** and **rating_denominator** columns were assessed for inconsistent values. The threshold for the numerator was fixed at 14 as most of the tweets had ratings between 1 and 14. Rows with numerators outside this threshold were dropped. Also the denominator was fixed at 10 and rows with denominators outside this threshold were also dropped. In the end the **rating_denominator** column was also dropped since all the denominators were now uniform.
5. The 4 dog stage columns: **doggo**, **floofer**, **pupper** and **puppo** were condensed or melted into one column **dog_category.** In this column every instance where any of the dog stage appeared was entered in the column and instances where none of the stages appeared was replaced with **none**.

6. A number of inconsistent “names” in the **name** column were replaced with none. It was important to do so as opposed to entering them as missing values so as not to distort our dataset when the need to drop missing values arise.

The Image Prediction dataset was largely untouched. However, some feature engineering was carried out on it where all Dog breeds in the **p1**, **p2**, and **p3** columns and their confidence levels were extracted such that the prediction with the highest confidence level as well as the corresponding True prediction (**p1_dog**, **p2_dog** and **p3_dog**) was used to populate a new columns called **breed** and **confidence.** These new 2 variables were merged with the tweets archive dataset.

The data from the JSON file was untouched. It was instead merged with the twitter archive dataset to get the retweet count and favorite count in the tweets in the archive. Tweet IDs that do not have a match in the JSON file were dropped.

The resulting expanded dataset was saved as a new file **twitter_archive_master.csv.**