Data Wrangling

Gathering

First I imported all the necessary libraries that would enable me read the various file formats given.

Using the imported libraries:

- I read the CSV file using pandas
- Using the request Library, I downloaded the Image predictions data from the URL given, created a TSV file and wrote the contents into the TSV file.
- Next I read the created TSV with pandas.
- I loaded API data stored in the 'tweet-json.txt file as a dataframe.

Assessing

The data for all 3 files were assessed visually and programmatically for quality and tidiness issues.

We find Issues such as wrong datatypes as well as redundant columns in the frist dataframe named df_twitter.The dataframe containing image predictions, named df_ url contains confusing column names ,the data extracted from the API has a id column which doesn't match with the tweet _Id column name we have in the other two dataframes.We also have missing data, wrong data values in the dataframes. Meanwhile viewing related data across three dataframes is tedious, so the three dataframes would need to be combined into one.

Cleaning

In cleaning, I started by merging all dataframes into one, after which I made a copy of the merged dataset. I then proceeded to compute all dog stages under one column 'dog_stage' hence the columns for each dog_stage were no longer needed, so I dropped them.

I also dropped the following columns because I wouldn't be needing them.Columns such as 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' thereby reducing my columns from 30 to 23.

I edited the column names to be better descriptive, also made changes to the values in the rating numerator and denominator columns, ensuring all denominator values are equal to 10.

Since we were told not to include retweets in our analysis, I dropped rows which were retweets
Next I changed the timestamp datatype, from object to datetime.