Wrangle Report

After merging the data we downloaded with the three columns from the data pulled from the twitter api 'id','retweet_count', 'favourites_count'. However, since there is a tweet_id column in the original data, adding another id column would be redundant, I only used the ‘id ‘column to join.

One of the data quality problems that exist in the data set is the presence of ‘None’ instead of NaN when a value is not available. To tackle this I used the replace function to replace all instances of ‘None’ with an empty string. This problem was in the dog stages columns (doggo, floofer, pupper, puppo) and the ‘name’ column. A data tidiness issue is having a separate column for each dog stage where it can be combined into one column named ‘dog_stage’. This will cause the rows that have two dog stages since, dog stages are not mutually exclusive due to the existence of multiple dogs in one photo, so this is another quality issue. To tackle this after merging the dog stages columns, I checked for rows that have multiple dog stages concatenated. Then I changed them to include a hyphen in between.

Since we don’t want tweets that are replies or retweets, we drop those rows that have a value in either of the columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, or retweeted_status_user_id. Then, drop the columns themselves as we have no use for them.

Another data quality issue is the timestamp column; it is a string and has a time zone component that is useless. I changed that to datetime and formatted the dates accordingly.

A glaring data quality issue is the inconsistent ratings with different numerators and denominators. This is caused by having multiple dogs in a picture. By calculating the number of dogs in a picture, then using this dog count to divide the numerator to get a correct rating. In addition to changing all denominators to 10 we can achieve consistent ratings.

The last data quality issue is having ‘a’ in the name column, I replaced that with an empty string to having consistent null values in all columns.

To summarize all the issues:

Data quality issues:

- Cells that contain none in the dog types columns.
- None in the column name.
- Drop rows that have a in_reply_to_status_id and retweeted_status_id value, these are retweets or reply tweets.
- The +0000 in the timestamp, it's the same for all rows and doesn't serve a purpose.
- Inconsistent rating numerators.
- Inconsistent rating dominators.

- Dog types are not mutually exclusive; each dog can have one or more dog stage.
- Names with 'a' in the column.

Data tidiness issues

- There is no need for four columns to represent dog stage; one column is enough to represent dog stage.
- Remove in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id columns after removing the rows that have values in them.