Following Quality and Tidiness Issues identified and resolved in this project's Data Wrangling

Quality Issues:

1. Missing Records

i) Coulmns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp having large numbers of NaN

Solution: Dropped the columns

ii) 2 tweet_ids not found in data obtained through api. 2356 ids are there in twitter-archive-enhanced.csv wheras tweet_json.txt contains 2354 unique tweet ids.

Solution: With merge of two data sets kept only the common tweet ids

iii) 281 tweet_ids are missing from image-predictions.tsv in comparison to twitter-archive-enhanced.csv

Solution: With merge of two data sets kept only the common tweet ids

iv) None and a do not seem to be names, 576 names are "None" and 55 names are "a"

Solution: No change, as the issue is not affecting any other analysis

2. Validity issues

i) More than 10 denominator in the rating_denominator column

Solution: Changed all values which are not 10 to 10

ii) more than 20 numerator in the rating_numerator column

Solution: Changed all values which are more than 20 to 20

iii) Only 307 tweets have dog stage classified correctly

Solution: Created a separate data set of these 307 values and analysed them for dog stage

iv) Rating 15 favorite_count is 0 which is perhaps indicating data quality issue

Solution: Dropped the row

3. Accuracy issues

i) Source column does not contain full source content and has curtailed source data which is not useful

Solution: Dropped the column

Tidyness issues:

i) Userids in in_reply_to_user_id, retweeted_status_user_id are being displayed in scientific notation

Solution: Changed the width of all columns to display entire User ID in original form