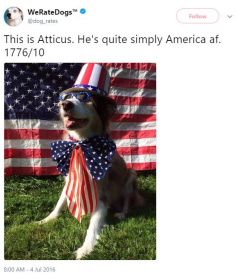
*Used GoogleSheets for visual assessment.

Twitter_archive - Quality

- missing some expanded_urls. This might be because there isn't an image with the tweet. The expanded url is housed under the media tag, no media, no expanded url.
    - Extracted the expanded_urls from the twitter api json. Some of the expanded urls in the twitter_archive didn't match the json expanded_url. These non-matching urls were either invalid urls or expired urls. The json url works, I elected to delete the 'expanded_urls' column.
- data contains retweets. Only want original tweets
    - Deleted tweets that had a non-null value for retweeted_status_id.
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, are floats, so they are written in scientific notation
    - The in_reply_to_status_id and in_reply_to_user_id still displayed in scientific notation after being converted to integers. To fix this issue, I extracted the string versions of the two tags from the Twitter api. Then I replaced in_reply_to_status_id and in_reply_to_user_id columns with the extracted columns. I elected to leave them as strings since we use the whole number as identification not for calculations.
- timestamps are objects
    - Changed into datetime object.
- not all the denominators are 10*
    - This information was extracted from the tweet text using the format rating_numerator / rating denominator. Some of the denominators weren't '10' because dates and other phrases use the same format. Others weren't 10 because it was a rating for a group. I was able to programmatically change them using the following formula:

```
#finding individual rating numerators
num_dogs = twitter_archive_clean1['rating_denominator']//10
twitter_archive_clean['rating_numerator'] = twitter_archive_clean['rating_numerator']//num_dogs
```

- Some of the numerators were below ten
    - Some of the lower ratings were due to incorrect extractions of a decimal numerator (i.e 13.5). Others were due to just extracting the first rating of multiple ratings. These ratings were replaced with the average of the ratings in the tweet.
- some numerators are really high (1776, 420, etc)*
    - Some of the high numerators were incorrectly extracted, but some of the high ones were done in theme:



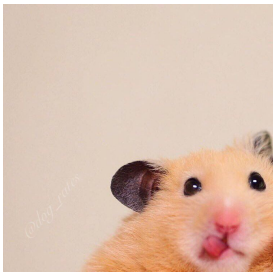
- some names are not actually names (a, an, etc)
    - I created a subset dataframe that included lowercase ‘names’ and the null values from the ‘name’ column. Some of the names had to be manually changed, but some of them could be extracted programmatically because they appeared after the word ‘named’

Tidyness-twitter_archive

- dog stages are in multiple columns.
    - Combined the dog stages into one column. Some of the tweets had multiple dog stages, so I double checked that there were actually multiple dogs. I then manually changed the stages for the single dogs from the multiple dog stages.

Quality - tweet_image_predictions

- 324 entries do not contain any dog breed predictions among the three choices, however some of them are images of dogs. We only want tweets that contain images of dogs.
    - I subset tweet_image_predictions to look at tweets without any dog breed predictions. I then deleted the tweets that didn’t have dog images. While I was sifting through the 324 images, I realized that the image predictor was good, but not great. So then I looked at tweets with at least one non-dog breed prediction and deleted tweets that weren’t dog predictions. Super tedious part, but at least the pictures were cute.



Quality - tweet_counts

- timestamp isn't a datetime object.
    - Converted to datetime

Tidyness - twitter_archive_master

- Duplicate columns: tweet_id, expanded_url, in_reply_to_status_id, in_reply_to_user_id, timestamp.
    - Kept the duplicates columns from tweet_counts (the data from the Twitter API). See the point about the missing expanded_urls.
- remove _y from in_reply_to_status_id, in_reply_to_user_id, timestamp columns
    - The ‘_y’ was put onto the column names to distinguish them from the same columns in twitter_archive.
- delete is_dog column. We know they are all dogs from previous cleaning.
    - This column was made to clean tweet_image_predictions, so it is not needed in the master dataframe.