wrangling efforts

in this project, I will be wrangling, analyzing and visualizing the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The data I will be gathering, assessing, cleaning, analyzing, and visualizing will come from a downloadable CSV file, a URL link, and a JSON file extracted from twitter's API.¶

A) Gathering Data

- gather the twitter archive data from the manually downloaded file
- gather the image predictions data from the provided url
- gather Each tweet's retweet and like counts from the JSON file Acquired from Twitter's API u sing tweepy

B) Assessing Data

Assessing & and later cleaning the data upon which the analyses and visualizations are based

Assessment documentation for twitter archive data

Quality:

1) there are 78 reply

2) there are 181 retweets

3) there are rating numerator above 20

4) there are rating denominator that are not 10

Assessment documentation for image predictions data

Quality:

5) there are 543 predictions (algorithm's #1) that are not for dogs

6) there are 887 predictions with low accuracy (<70%)

Tiredness:

1) observations are the same type as the twitter archive data

Assessment documentation for retweet and like data

Quality:

7) the key column is named 'id' which is different than the key column name in the other 2 tables 'tweet_id'

8) the column lang contains an abbreviation of the language that is not clear

Tiredness:

2) observations are the same type as the twitter archive data & the image predictions data

C) Cleaning Data

Assessing & and later cleaning the data upon which the analyses and visualizations are based

- Quality issue 1 in twitter archive data: 78 replies need to be deleted as these are not original ratings
- Quality issue 2 in twitter archive data: 181 retweets need to be deleted as these are not origional rat ings
- Quality issue 3 in twitter archive data: rating numerator above 20 are typos and needs to be deleted
- Quality issue 4 in twitter archive data: rating denominator that are not 10 are typos and needs to be deleted
- Quality issue 5 in image predictions data: 543 predictions(algorithm's #1) that are not for dogs need s to be deleted as these are false predictions
- Quality issue 6 in image predictions data: 887 predictions with low confidence (<70%) needs to be d eleted to prevent it from affecting the analysis findings
- Quality issue 7 in retweet and like data: the key column is named 'id' which is different than the key column name in the other 2 tables 'tweet_id', it has to be changed to ensure successful merging l ater
- Quality issue 8 in retweet and like data: the lang column has only an unclear abbreviation, this need s to be replaced with the language name
- Tiredness issue 1 in image predictions data: observations are the same type as the twitter archive d ata, so I will join both data frames
- Tiredness issue 2 in retweet and like data¶: observations are the same type as the twitter_archive_ master, so I will join both data frames