Report

Gathering phase

There were three main sources for the data to deal with. The three pieces of data were obtained and represented as pandas dataframes in this phase:

- The WeRateDogs Twitter archive (downloaded 'twitter-archiveenhanced.csv' manually)
- The image predictions from Twitter ('image-predictions.tsv'). This document was downloaded.
- file named 'tweet json.txt.' was downloaded due to restriction of API

Assessing Phase

The visual evaluation was completed in a Jupiter notebook. Followed by programmatic assessments, and then untidy structures were fixed to help with the rest of the quality concerns.

Codes used for assessment

df.sample(20)

df.info()

df.describe()

Quality

-change datatype for the tweet_id to string since we won't do calculation using them

-drop retweets rows.

-drop column related to retweets or replies info like retweeted_status_id, retweeted_status_user_id,

tweet_id, in_reply_to_status_id, in_reply_to_user_id, and retweeted_status_timestamp.

-change timestamp into time.

-remove invalid names with lowercase

-drop rows without dog_stage

-change data type of rating_numerator and rating_denominator-rating_numerator should be a float

-change rating_denominator that are less that 10

Tidiness

1- There are too many datasets and their overall structure is untidy.

2- join the columns of doggo, floofer, pupper, and puppo into dog_type.

Cleaning phase

This process starts by copying data into new dfs then making changes into these new dfs

Quality

- timestamp is string and is changed into datetime using pd.to_datetim()
- drop column related to retweets or replies info like retweeted_status_id, retweeted_status_user_id, tweet_id, in_reply_to_status_id, in_reply_to_user_id, and retweeted_status_timestamp. Using .drop() function.
- remove retweets rows using df1_clean=df1_clean[pd.isnull(df1_clean['retweeted_status_user_id'])] by not copying the row if there is a value for 'retweeted_status_user_id'
- dropping rows with lowercase names using drop() and lowercase function to find the indices
- drop rows without dog_type using drop.()
- replace demoninators that are not equal to 10 with 10 using loc()
- change data type of rating_numerator and rating_denominator-rating_numerator should be a float using astype(float).
- change datatype for the tweet_id to string since we will not do calculation using them through astype(str).

Tidiness

1- A single dataset was created using merge()
2- join the columns of doggo, floofer, pupper, and puppo into dog_stage after replacing none with nan and joining the columns using + operator.