Udacity Data Analyst Nanodegree Program

wrangle and analyze data

Gathering data:

- twitter-archive-enhanced.csv is read by pandas using read_csv.
- image_predictions.tsv is downloaded by requests using gets then read by pandas using read_csv.
- tweet_json.txt is first writing by using the API then opened line by line in python and finally read in pandas by making a DataFrame.

Assessing data:

Quality:

twitter-archive-enhanced.csv:

- “timestamp” is string
- there are 181 retweets in the dataframe
- “in_reply_to_status_id”, “in_reply_to_user_id”, “retweeted_status_id” and “retweeted_status_user_id” are float64
- “retweeted_status_timestamp” is a string
- “expanded_urls” has nulls
- "None"(string) as a name
- remove the “a” tags form “source”
- “rating_denominator” has values other than 10.

image_predictions.tsv:

- missing data (2075 for 2536)
- 3 predictions per row
- no need for the field “img_num”

tweet_json.txt:

- missing data (2354 for 2536)
- “Tweet_ID’ is a String

Tidiness:

twitter-archive-enhanced.csv:

- “doggo”, “floofer”, “pupper” & “puppo” are "stage"s of a dog
- no need for “retweeted_status_id”,”retweeted_status_user_id” and “retweeted_status_timestamp”
- ratings should be one field

image_predictions.tsv:

- None

tweet_json.txt:

- None

Cleaning data:

Fixing Tidiness:

twitter-archive-enhanced.csv:

- drop “retweeted_status_id”,”retweeted_status_user_id” and “retweeted_status_timestamp” fields (the rows that were retweets where also dropped)
- combine “doggo”, “floofer”, “pupper” & “puppo” fields into one categorical called “stage”
- make a “rating” field by dividing “rating_numerator” by the “rating_denominator” field (if the “rating_denominator” = 0 then the “rating” would be set to 1) then multiplied 10 (this fixes “rating_denominator” not being 10 problem)

image_predictions.tsv:

- created “isDog”, “prediction” and “confidence” fields which takes one of the predictions which has both a high confidence coefficient and said prediction is a dog.(if none are found the values say “No prediction")
- the field “img_num” was dropped

tweet_json.txt:

- the field “Tweet_ID” was renamed to “tweet_id”

Fixing Quality:

twitter-archive-enhanced.csv:

- the field “timestamp “ type was changed to datetime64[ns]
- the field “in_reply_to_status_id” and ” in_reply_to_user_id “ types were changed to String
- the rows that had NaN in the “expanded_urls” field were dropped
- the name “None” was changed to NaN in the “name” field
- the “a” tags were removed using BeautifulSoup in the “source” field

image_predictions.tsv:

- None

tweet_json.txt:

- the field “tweet_id” type was chaged to int64

join all data into one csv:

using the .merge method in pandas I was able to merge all the data frames into one data frame and save it as "twitter_archive_master.csv"