Report for Data gathering, assessing, cleaning, and storing parts of twitter project

Data gathering:

- **api_df:**
  - Collected data from twitter archive using twitter API and saved the data json information in a text file.
  - Looped over the data text file line by line, read each line using json.loads() function, extracted the variables I'm interested in (tweet_id, retweet_count, favorite_count,user_followers_count) and saved them with their values as key value pairs in a dictionary object. All dictionary objects were stored in an array. The final result is a list of dictionaries where each dictionary represents a tweet.
  - Converted the list of dictionaries into a dataframe using pandas DataFrame function and named it api_df.
  - Note: my twitter developer account wasn't activated by the time I completed this project so I used the code and text file provided by udacity mentors.
- **archive_df:**
  - read twitter-archive-enhanced.csv file into a dataframe named archive_df.
- **image_predictions_df:**
  - read image-predictions.tsv file into a datarame named image_predictions_df specifying the separator as tab separator (\t).

Data assessment:

- **api_df:**
  - Displayed the table head and info to inspect variables datatypes and values.
  - **Quality Issues:**
    - tweet_id should be string not int
- **archive_df:**

- Displayed tab head and info to inspect variable datatypes and values to find quality and tidiness issues.
- Displayed values of certain variables as name, source, and rating denominator to further inspect them and find null and wrong values.
- **Quality Issues:**
  - some names are uppercase and some are lowercase.
  - some dogs have none as names.
  - tweet_id should be string not int.
  - source column contains whole html element instead of source name only.
  - tweets with expanded_urls value of null should be removed.
  - rating_numerator, rating_denominator are int when they should be floats.
  - timestamp is string not datetime.
  - retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp all mostly nulls so should be removed.
  - in_reply_to_status_id, in_reply_to_user_id are mostly nulls so should be removed.
- **Tidiness Issues:**
  - four dogs columns should be combined into one.

- **image_predictions_df:**
  - Displayed tab head and info to inspect variable datatypes and values to find quality and tidiness issues.
  - **Quality Issues:**
    - tweet_id should be string not int.
    - rename jpg_url column name to image_url
- **Tidiness Issue:** api_df, archive_df, image_predictions_df should all be merged into one master table named twitter_archive_master.

Data cleaning:

- Followed the define-code-test pattern to define my solution and code it then test the output to check if it aligns with my solution definition or not.

- **archive_df:**
  - **Quality Issues:**
    - Changed tweet_id from int datatype to string to eliminate any accidental ordering or misunderstanding that any order is implied through the id.
    - make all dog names lowercase
    - change all dog names that are none to np.nan.
    - change all denominator values to 10.
    - make timestamp datatype datetime instead of string
    - make source column contain only the source name instead of the whole html element.
    - remove retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp,in_reply_to_status_id, in_reply_to_user_id columns
    - drop tweets with null expanded_urls values.
    - changed rating_denominator and rating_numerator datatypes from int to float.
  - **Tidiness Issues:**
    - combine 4 dogs columns [doggo,floofer,pupper,puppo] into one column named dog_stage

- **image_predictions_df:**
  - **Quality Issues:**
    - Changed tweet_id from int.
    - Renamed jpg_url column to image_url to describe better the content of the column (url of the image being classified by the network).
- **api_df:**
  - **Quality Issues:**

        - Changed tweet_id from int.

Data storing:

- **twitter_archive_master_df:**
    - Stored master table into csv file named twitter_archive_master.csv.
- **archive_df:**
    - Stored clean version of archive_df table into csv file named twitter_archive_clean.csv.
- **image_predictions_df:**
    - Stored clean version of image_predictions_df table into csv file named image_predictions_clean.csv.