Data Wrangle Report

1. Gathering Data:
   1.1. Downloaded the csv file(twitter-archive-enhanced.csv) from udacity manually and imported it into a DataFrame as “df”.
   1.2. Used the requests function to get the tsv file from given url, and write the file in a csv file through csv.writer() function. Then imported it into a DataFrame as “df_image”.
   1.3. Used the library tweepy and its function .get_status to access the Twitter API, then converted it to a Python object into the “tweet_json.txt” as a JSON formatted data with json.dump(). (The accessing time is about 2000 seconds) Read each JSON file with json.loads() and extracted the attribute tweet_id, retweet_count and favorite_count to create a DataFrame as “df_api”.
2. Accessing Data:
   2.1. Visual assessment: Each DataFrame was shown in the Jupyter Notebook.
   2.2. Programmatic assessment: Used .shape, .sample(), .info(), .describe(), etc. to look over.

- Used df.info() and found out there are some missing data

-Missing data on in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls columns

- Used df.info(), df_image.info(), df_api.info() and found out some wrong datatypes.

-Erroneous datatypes(tweet_id, timestamp columns, floofer)

- Used df.tail() and found out that some invalid names in name column.

-Some wrong name (a, just) in name column

- Used df.describe() and found out some rating denominators were not 10 which was not consistency and there were some rating numerators were extreme big which may be outliners.

-Some numbers are not 10 in rating_denominator column

-Some rating numerator are extreme big(outlier)

- Used df.source.unique() and found out there were only 4 unique values. The tag in these 4 values were same.

-Too many useless information on source column

- Used df[(df['doggo'] == 'doggo') & (df['pupper'] == 'pupper')] and found out some pictures had two dogs. However, tweet_id 817777686764523521 just had one and was mislabeled ‘doggo’ because its Instagram name is didodoggo.

-tweet_id 817777686764523521 not have doggo this attribute

- Doggo, Pupper and Puppo can be combined to one column.
- Select the useful data from df_image and df_api to df and delete some useless column in df.

-doggo, pupper, puppo should be combined to one column
-retweet_count, favorite_count in df_api table should be the part of `df` table
-text column does not have useful function in this project
-the final prediction for the image should be the part of `df` table

3. Cleaning Data:

Each unclean data is documented in Jupyter Notebook with ‘Define’, ‘Code’, ‘Test’.

4. Storing and Acting on Wrangled Data

Store the final DataFrame with df_clean.to_csv('twitter_archive_master.csv').