Wrangle report

Data gathering-

1. The first step of gathering is to upload twitter-archive-enhanced.csv to a dataframe called ‘enhanced’
2. The next step is to download 'image_predictions. tv and open it.
3. In this step I was trying to work with API and I couldn’t pass the authentication, so I download 'tweet_json.txt' from udacity.
4. I upload 'image_predictions.tsv' and 'tweet_json.txt' to dataframe image_pre and re_fav.
5. I merge all 3 dataframe in an inner join because I want to have only tweets with images.

Assessing data-

1. In the visual assessment that I did in Excel, I found the issues:
    - p1, p2, p3 and name sometimes lowercase and sometimes title
    - in tweets that have 2 images or more in expanded_urls columns Have the same link several times.
    - doggo, floofer, pupper, and puppo change them to one column called 'stage'
    - rating_numerator, rating_denominator change to 1 column.
    - remove replay and retweets rows
    - drop 5 blank columns after 7 ['in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp']
2. With info() I found these issues:
   Incorrect datatype in columns tweet_id, timestamp, and stage
3. with values counts I found these issues:
    - Nulls represented as None in doggo, floofer, pupper, puppo, and name.
    - a in the name is equal to null, I saw 55 ‘a’ and after a little check, I found that the name has been taken from the text. when it writes this is [dog name], and a lot of times people tweet like: this is a very cute dog, and it thinks that ‘a’ is the dog name.
4. from describe() I founds:
   rating_numerator and rating_denominator are not always on the same scale.
5. I decided that if I merge to 1 column called ‘stage’ the stage needs to be a categorical datatype.

Cleaning Data

1. First I copy the dataframe
2. I fixed the tidiness issues before I merge doggo, floofer, pupper, and puppo I replace None with null. I sorted the value and drop duplicated with tweet_id. Because the values were sorted if the tweet id has a value in the stage it drops all the null values. And tweet_id doesn’t have value it keeps 1 null row.
   I divide rating_numerator with rating_denominator and multiply by 100 (%) to create 1 column called rating(%).

3. I fixed the quality issues, and change the type to all the columns that need astype and to_datetime.
   With str. replace I replace all None values to null, and ‘a’ to null.
   I change p1, p2, p3, and the name to lowercase.
   With the split, I keep only 1 link in expanded_urls.
   I filter the df that retweeted_status_id and in_reply_to_status_id will be with a null value,
   And with the info, I test and see that all 5 columns have only null values.
   And after that, I drop the 5 columns.

   After the cleaning, I imported the clean dataframe to CSV call ‘twitter_archive_master.csv’.