Reporting: wrangle_report

INTRODUCTION: "WeRateDogs" Twitter archive contains basic tweet data for all 5000+ of their tweets which have been filtered and enclosed in twitter_wnhance_archive csv file forming the basis of the analysis.

AIM OF THE PROJECT: The goal is to gather, assess,clean,analyze and visualize the dataset.

GATHERING DATA: We need to gather three different datasets enclosed in three different formats.

- The first one is a .csv file and can be download directly and read into pandas dataframe using .read_csv() function.
- The second one is a .tsv file and should be downloaded by importing "requests" library.
- The third one is a .json file and can be downloaded either by creating a Twitter account to get API credentials or directly by loading the "tweet_json.txt" file.

ASSESSING DATA: After gathering data, we need to do both visual assessment and programmatic assessment.
Here are my observations from the visual assessment and info() from gather section:

- df_twitter_archive: This dataframe is not tidy, specifically with the dog stages.We can group all those columns into a single column and drop the rest of columns to make it tidy.There are missing values in in_reply_to_status_id ,in_reply_to_user_id, retweeted columns,Nan in name column and change the tweet_id datatype to object and timestamp datatype to datetime.
- df_image_predictions: The types of dogs in columns p1, p2, and p3 had some uppercase and lowercase letters.
- df_tweet_data: This doesn't require a separate table and has to be joined on 'tweet_id'.

Here are my observations from the programmatic assessment:

Quality issues

Twitter_Archive

1. There are 181 retweets and 78 reply tweets which need to be removed as the key points focus on original tweet ratings and no "retweets"; this data is stored in the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id ,in_reply_to_user_id.
2. Missing Nan values(too many names to verify and will not be cleaned) and 109 names having a,an,Bo are like part of the name which are not real names and need to be replaced with none.
3. The max of 'rating_numerator' is so large and so we have to clean the data with rating numerator <=15.
4. The value of 'rating_denominator'are greater than 10 which is not possible ,so we have to clean the data with rating_denominator not equal to 10

5. There are only 4 types of values in the source column, and they can be simplified by using the display string portion just before the final "<\a>".
   Image_predictions
6. We have to change the datatype of tweet_id column to object instead of int64.

Tidiness issues

1. In image_predictions dataframe,we need to remove 324 non dog breeds and introduce two new columns for breed and confidence to merge with the twitter_archive dataframe.

CLEANING DATA: Once we have assessed our data and documented them as above, we need to clean the data by creating a copy of our datasets and then using various pandas functions. Some of them which I have used are .astype(),str.replace(),.to_datetime(),pd.merge(),.drop(), using advanced indexing etc.

STORING DATA: After cleaning data,in my case, 8 Quality issues and 2 Tidiness issues we need to store wrangled datasets to a CSV file named "twitter_archive_master.csv" using .to_csv() function.