Wrangle_report

Here in this project of data wrangling we have three dataset collected from different sources. These three dataset files are as follows:

1. Twitter_archive_enhanced.csv
2. Tweet_json.txt
3. Image_predictions.tsv

Let's discuss about issues and challenges I faced with these files and procedure to rectify the same in details.

1. Twitter_archive_enhanced.csv:

The largest among the three datasets, provides all basic data for the analysis and insights. So, our focus should be to wrangle it properly. The dataset contains 17 columns and 2356 rows. By seeing, *retweeted_status_id* column we can easily sense that some re-tweeted data are present which we do not need for our analysis. After removing re-tweet data we are left with 2173 records. This can be achieved using code *df_archive.text.str[:2]!='RT'. Text* column start with **RT** for re-tweets.

Through inspection I found out that one tweet with tweet_id *835246439529840640* has *rating_denominator* as 0 which seems incorrect. So, I verified this with tweet 'text' using *print(df_archive_clean['text'][313])* and found that actual *rating_denominator* is 10. Also, there are some extraneous columns like *in_reply_to_status_id, in_reply_to_user_id'*, *retweeted_status_id,retweeted_status_user_id* and *retweeted_status_timestamp* that can be removed safely. For this, a simple *drop* statement will do the job.

Timestamp is stored as string variable, However it should be a *datetime* data type. This data type conversion can be easily achieved by pandas ***to_datetime*** function. One more variable ***Source*** has some extraneous characters that need to be removed for better analysis and insights. Column ***text*** has many valuables information such as rating, dog name, dog stage and short-url which could be extracted to separate columns.

Using *text* column another useful information *dog name* can extracted. As there is *name* column already exists which contains *dog name*, but some extraneous data is included in it. So, my task is to verify the correct ones and remove those not real dog names like 'a',' an', 'the' etc. To do this I used regular expression like this:

```
keywords=['named','name is','name to','This is',
                'THIS IS','Meet','hello to','this is','RIP','featuring','NAME. IS.']

def dog_name(txt):
for key in keywords:
    if key in txt:
      return  re.findall(r'{} (\S+)'.format(key), txt)[0]
```

Two important columns are *rating_numerator* and *rating_denominator.* However, some of *rating_numerators* have fraction values which are not extracted correctly. To extract it correctly there need of some change in regular expression, I used *r'([0-9]+[0-9.]*/[0-9]+[0-9]*)'* that works fine. Now these two can be used to create a new column *rating* (=rating_numerator/rating_denominator) which will help in compare tweets magnitude wise.

One more important issue is that columns *doggo, floofer, pupper and puppo* are in separate columns, however, there there is no need of this and for better table structure these data can be put into one single column ***dog_stage.***

2. Tweet_json.txt:

The smallest one among all three, it contains *tweet_id, favourite_count* and *retweet_count.* Only one major issue with dataset is that data type is string and needs to be change to integer. I can use *astype* for this task. Also, there is no need to maintain a separate table for these two columns *favourite_count and retweet_count* and better be merged with *twitter_archive_enhanced.*
This task can be done as follows:
pd.merge(df_archive_clean,df_tweet_clean,on='tweet_id',how='left').

3. Image_predictions.tsv:

This dataset contains some of the very important information. It contains prediction's algorithm output of the dog images. The algorithm predicts dog's breed with percentage of confidence. However, there are output of three algorithms is given in dataset. To make the analysis simple I will use only best prediction algorithm result i.e. p1, p1_conf and p1_dog.
This dataset should also be merged with the twitter_archive_enhanced.

Finally, after doing above cleaning and merging I exported the data in CSV format file named: ***twitter_archive_master.csv.***