Data Wrangling Report

Investigate twitter_archive_enhanced.csv

The dataframe “tae” is created from twitter_archive_enhanced.csv which is merged data from twitter_archive_enhanced.csv, tweet_json.txt .

tweet_json.txt , twitter_archive_enhanced.csv , image_prediction.tsv are first visually analyzed by opening as csv file on google sheets.

Dataframe tae is investigated programmatically using below mentioned panda methods.

Tae.info() displays

- in_reply_to_status_id and in_reply_to_user_id are floats.
- Timestamp is an object.
- There are 181 values in retweet_status_id and retweet_status_user_id columns.
- The columns retweeted_status_id and retweeted_status_user_id exist which are not needed for analysis.

Tae.describe() displays numeric columns. tae.tail(),tae.sample().tae.info() used to take the first look at all the columns info and sample values.
Min max rating denominator and min max rating numerator are evaluated
To help understand range on ratings. Ratings are needed to be calculated in further wrangling efforts. The value counts in ratings numerator and denominator are analyzed for finding the range. The Zero value in numerator and denominator was not detected in min max analysis. The value_counts showed there are few records with 0 value. These rows will need attention.

Below quality issues are detected in twitter_archive_enhanced.csv

Issue 1. Retweet id is present for many rows

Issue 2. Ratings are accessed by programmatic addition of column ratings and fixing the rows with 0 denominator values

Issue 3. 181 are retweet

Issue 4. in_reply_to_status_id is float and has nulls

Issue 5. in_reply_to_user_id is float and has nulls

Issue 6. retweeted_status_id,retweeted_status_user_id are not needed for analysis.

Issue 7. Names with value 'None' or 'a'

Issue 8. tweet_id is defined as object and int64

Issue 9. timestamp is object

Issue 10 expanded_url is null

Investigate image_prediction

Describe(),info(),value_counts(),sample(),head(),tail() are used for general programmatic assessment of columns,data type and values.
1.P1_conf > p2_conf > p3_conf shows that p1 is most predictable value

Investigate jasondf created from tweet_json.txt

Describe(),info(),value_counts(),sample(),head(),tail() are used for general programmatic assessment of columns,data type and values.
1. tweet_id is defined as object and int64 in twitter_archive_enhanced.csv & image prediction
2. retweet_count & favorite_count are objects in jason df.

Below tidiness issues are noted from investigating above 3 dataframes:
1.Columns doggo ,floofer, pupper, puppo need to be under one column heading Doggo Lingo
2.Column retweet_count , favorite_count of jtweet_jason_df need to be with columns of twitter_archive_enhanced.csv as it makes more relevance along with other variables
3. Image and highest prediction need to be with columns of twitter_archive_enhanced.csv as it makes more relevance along with other variables

Cleaning Efforts:

- Drop rows with retweeted_status_id <> 0
- Remove null from in_reply_to_status_id and convert to int
- Remove null from in_reply_to_user_id and convert to int
- Convert timestamp from object to datetime
- Change name = ‘none’ and ‘a’ to NoNameDogs
- Drop row with rating denominator min = 0 and numerator min = 0. This decision is made to avoid math error issue for calculating ratings_round = numerator/denominator
- Calculate and Round the ratings: A New column is added ratings_round to store the calculated ratings.
- Replace nulls and combine the columns doggo,floofer,pupper,puppo

- Fix the data type of jasondf columns and merge columns retweet_count , favorite_count of jasondf with twitter_archive_enhanced on tweet_id as it makes more relevance along with other variables.
- join image_prediction with tae_merge for image_url and p1