WeRateDogs - Twitter Data

DATA GATHERING

Instructions were given by the udacity instructor on how to proceed in gathering data.

- I downloaded the data, which is a given CSV file and named as twitter-archive-enhanced.csv.
- Next, I programmatically downloaded the file image predictions file, which is in the .tsv format.
- Then, I downloaded 'tweet-json.txt' from the udacity platform as I was having issues confirming my twitter developer account. I read the API pseudo-code and understood it before proceeding to the next step

DATA ASSESSMENT

Each table was displayed in its entirety by displaying the pandas DataFrame that it was gathered into. This task is the mechanical part of the visual assessment in pandas. Steps taken while assessing dataframes include:

- The first five rows of the dataframe were viewed to see if any anomaly such as column names and misspelling could be seen easily.
- Then null values were checked
- Duplicate rows were also investigated.
- The numerical values were then described to check for outliers and weird values.
- Then the info of each column was investigated.
- Lastly, we checked the datatypes for irregularities.
- Then based on some view observations, various columns were investigated

The columns were well explained https://sfm.readthedocs.io/en/1.4.3/data_dictionary.html . for better understanding of the datasets

Quality

I. Missing values in some columns from archive_df
II. outrageous and inconsistent values in rating numerator and denominator
III. one rating has a zero denominator
IV. weird names found for dogs - 'infuriating', 'just', 'life', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'
V. timestamp and retweeted_status_timestamp must be of datetime instead of the object

VI. comparing both image_prediction_df and tweets_info_df to twitter_archive_df we see that they both dont have complete tweeter id like twitter_archive_df The columns which have missing values in doggo, floofer, pupper , puppo are written as None instead of NaN

VII. non discriptive columns headers(p1,p1_conf,p1_dog,p2,p2_conf,p2_dog)

Tidiness

i. Plenty of columns explaining dog stage in archive_df when they could easily be merged as one

ii. had to merge the three dataset to get enough details

iii. Removal of retweets and replies as not all tweets were dog ratings

DATA CLEANING

The following steps were carried out when cleaning the data;

- All three datas were copied to a different dataframe so a not to deal or mess with the first one
- Row with zero”0” value ranking denominator was removed
- Timestamp and retweeted_status_timestamp were converted to datetime datatype
- Renaming columns dealing with prediction in the image_prediction_df to a more self-explanatory name
- None value in the dog stage columns was converted to an empty string for easy concatenation
- Renaming variables in source column for easy understanding