Wrangle Report

June 24, 2020

1. Gather Data

Data is gathered from 3 resources like below

- From given file 'twitter_archive_enhanced.csv' and set as ***twitter_arc*** dataframe
- Extract data programmatically from this URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv set as ***images*** dataframe
- Extract data from twitter api using tweepy library and saved data in JSON format in 'tweet_json.txt' file set as ***twitter_api*** dataframe

2. Access

Using panda library method info() get all three of dataframe detail and check duplicate data available or not and check data type of column and which column data contain NaN value and which try to find out which column is need to merge and set as one column.

3. Clean

twitter_arc dataframe

- Set tweet_id as integer in all dataframe and convert timestamp
- Set tweet_id is an integer
- Convert timestamp and retweeted_status_timestamp which is currently of type 'object'
- Name has values that are the string "None" instead of NaN
- Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN)
- doggo, floofer, pupper, and puppo have values that are the string "None" instead of NaN
- Incorrected ratings on rating_numerator and rating_denominator.

twitter_api dataframe

- There are 11 missing tweets compared to the twitter_arc datagrame (I am assuming they have been deleted)

images dataframe

- There are 2356 tweets in the twitter_arc dataframe and 2075 rows in the images dataframe. This could mean that there is missing data, or that not all 2356 of the tweets had pictures.
- tweet_id is an integer
- p1, p2, and p3 contain underscores instead of spaces in the labels

4. Tidiness Issues

twitter_arc dataframe

- 1 variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

twitter_api dataframe

- twitter_api data should be combined with the twitter_arc data since they are information about the same tweet

images dataframe

- images data could be combined with the twitter_arc data as well since it is all information about 1 tweet

5. Store Data

Store data in new csv file as 'twitter_archive_master.csv' and analyze from that.