wrangle_report

October 17, 2020

0.1 Project Details

0.1.1 Data wrangling, which consists of:

Gathering data

- The WeRateDogs Twitter archive : downloaded manually
- The tweet image predictions : downloaded programmatically
- Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing data

- visualize, Detect and document

Quality issues twitter_archive_df

- only want original ratings (no retweets) that have images.
- delete unnecessary columns 'in_reply_to_status_id','in_reply_to_user_id','source','retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp','expanded_url'
- incorrect data types, change'timestamp' to datetime instead of object
- 'rating_numerator', and 'rating_denominator' have invalid values as '1','2'
- Name contain 'None', 'a', 'an', 'the'

image_predictions_df

- delete img_num column
- delete rows with null jpg_url
- delete rows with duplicated jpg_url

tweets_df

- rename id in tweets_clean to tweet_id
- delete unnecessary columns

Messy (Tidiness Issues)

- merge favorite_count and retweet_count to twitter_archive
- dog Stages in twitter_archive_df should be in one column as it is 1 variable
- predictions (p1,p2,p3) in image_predictions_df should be in one column
- p_conf (1,2,3) in image_predictions_df should be in one column
- Merge 'image_predictions_df'into 'twitter_archive_df'.

Cleaning data

- solve every issue by defining it, then coding the solution , and finally test it.

Storing, analyzing, and visualizing your wrangled data

- Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv.

- analyze and visualize 3 points

- . the most dog stage

- . the most prediction

- . the mean of favourite count grouping with dog stages

```
In [ ]:
```