WeRateDogsTwitterData
WranglingReport
Abdu Magdy Qnauy

Sources

I gathered data from agiven CSV,a website,and Twitter's API. I used tweepy to access the API and gather the JSON data for the tweets. Is to read the JSON data in a text file, then loaded what I needed into a pandas dataframe.

Process

I audited the data by checking datatypes,valuecounts, number of non-null entries,and numeric summaries. Data for some a few tweets. I combined(that is,inner joined) all three tables because each column is a feature of its tweet.I reshaped the dog_stages(e.g.,floofer,puppo,pupper,doggo) into a single column rather than multiple columns. I converted several columns to new datatypes:

- @archive_df timestamp column its dtype converted from string to datetime
- @archive_df name column its missing values converted from None to NaN
- @archive_df doggo column its missing values converted from None to NaN
- @archive_df floofer column its missing values converted from None to NaN
- @archive_df pupper column its missing values converted from None to NaN
- @archive_df puppo column its missing values converted from None to NaN
- @archive_df text column is cuted off with ellipses
- @tweets_df tweet_id column needs to convert its dtype from string to int

For each fixed issue,I identified the issue, stated my intention,then tested to ensure that I enacted my intention.

Storage

I stored the data in two CSVs, one for the tweet data and one for the image predictions.

Remaining Issues

- I noticed a strange interaction where a tweet's text was cutoff in the dataframe.
- I decided not to change the numerators or denominators for the dogratings because they probably were not mistakes.
- Some of the variables did not seem necessary or useful(e.g.,*in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_user_id*).