wrangle_report

August 30, 2022

0.1 Reporting: wragle_report

This report documents the steps taken in wrangling the WeRateDogs data. The wrangling was divided into 3 main stages namely; 1. Data gathering 2. Data Assessment 3. Data cleaning

0.1.1 Data Gathering

I acquired the data from three different sources; 1. The first data was twitter_archive_enhanced.csv which I manually downloaded from Udacity servers, then I upload it and read the data into a pandas DataFrame. 2. The second data was image_predictions.tsv was also hosted on Udacity's server. I downloaded it programmatically using the Requests library. 3. The last data was acquired from Twitter API. I used the tweet IDs from the WeRateDogs twitter_archive_enhanced.csv, I then queried 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.

0.1.2 Data Assessment and Cleaning

After gathering all three pieces of data, I assess them visually and programmatically for quality and tidiness issues and cleaned up the data. This included;

Quality issues

df_archive

- Droping columns thats has a lot of null values, or that are irrelevant to the data analysis such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp and retweeted_status_user_id.
- Changing rating for twitter ids that were erroneously captured eg chnage rating of twitter id 835246439529840000 from 960/0 to 13/10, twitter id 775096608509886000 from 9/111 to 14/10, twitter id 716439118184652000 from 50/50 to 11/10
- Droping rows that have a denominator that isn't 10.
- Changing timestamp data into datetime data type.
- Removing rows with rating_numerator above 20
- Removing rows in with missing values if needed

image-predictions

- Droping the img_num column because it will not be useful in the analysis

Tidiness issues

- The four columns doggo, floofer, pupper, puppo were converted to one column named dog_stage
- finally, I merged all the datasets to form one master data named df.

The cleaned data was then stored as a CSV file named "twitter_archive_master.csv"

In [ ]: