wrangle_report

July 25, 2022

0.1 Data wragling Report

In this report I will describe the wrangling efforts to assemble and clean the data from the WeRateDogs Twitter Archive.

0.1.1 Data Gathering

The data I gathered was obtained from 3 sources: 1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers. 2. The image predictions file, programmatically downloaded from the Udacity servers. 3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library

0.1.2 Assessing & Cleaning

I began the assessment by viewing the information on the archive table first, identifying several quality and tidiness issues. Then I went ahead to programatically access df_image_predictionsdata.

I identified 8 quality issues and 2 tidiness issues. To begin my cleaning efforts, I converted ["doggo", "flooter", "pupper", "puppo"] columns into one "stage" column, then droped the four columns.

I Converted the timestamp column to datetime data type.

I changed the odd dog names like "a" and "an" in the name column to 'none'.

As per the requirements, all rows containing non-null values in the retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp , and also in the in_reply_to_status_id and in_reply_to_user_id columns were dropped.

I dropped the "source column because I did not need it for my analysis

The rating_numerator and rating_denominator columns were checked for value ranges; I decided to keep only tweets with single ratings. Several tweets' ratings were manually corrected with values from the text. Tweets with large numerators were dropped, as the text didn't contain a valid rating (# out of 10). After the ratings were fixed, I dropped the rating_denominator column (it contained only '10's).

The 4 dog stage columns were melted into the stage column; tweets without stages were set to 'none'. Several had 2 stages set, so I kept only the one with the lower overall count.

The remaining cleaned columns in the archive table were reordered, then the table was saved to the new "twitter_archive_master.csv" file. The predictions file saved as image_pred_wrangled.csv