Wrangle Report

Introduction

This project covers the data wrangling process of a tweet archive of a twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. This report talks about a brief description of the techniques used in gathering the data, assessing the data, and cleaning the data. Three steps were carried out in the data wrangling process namely: gathering data, assessing data, and cleaning data.

Gathering Data

In the data gathering process, 3 files were gathered for the analysis from WeRateDogs **WeRateDogs twitter archive**, which is a csv file that was downloaded manually and contains some basic tweet data of 2300+ tweets.
The tweets Image predictions, which is present in each tweet according to a neural network.
Each tweet's retweet count and favorite (like) counts, which is a text file called tweet_json.txtthat contains the JSON data for each tweet's retweet counts and favorite counts
These 3 files were then loaded into a pandas dataframe from data assessment

Assessing Data

In the assessment process, I evaluated the 3 dataframes visually and programmatically. I assessed the dataframes for quality issues and tidiness issues.
The following quality and tidiness issues were observed:

Quality Issues

Twitter archive table

- change the timestamp datatype to datetime datatype and the tweet_id datatype to string datatype
- some names in the name columns are invalid names (like a, an, None, etc
- remove retweets by dropping columns with any retweet status like retweeted_status_id, retweeted_status_user_id,
- remove 'html' from source column for easy understanding
- rating_denominator and rating_numerator have invalid values i.e very high values that are considered
- drop columns that would not be needed for the analysis

Image predictions table

- drop img_num column
- duplicate enteries in the jpg_url column
- change tweet_id datatype to string datatype

Twitter Json (API) table

- change tweet_id datatype to string
- remove 'html' from source column
- remove 'retweeted' column

Tidiness Issues

- combine the 4 dog stage columns into one single column named dog class
- merge the 3 tables into one single dataframe

Cleaning data

The quality and tidiness issues identified were cleaned programmatically, like;

- dropping unnecessary or unneeded columns
- correcting invalid or inaccurate values
- correcting datatypes
- removing duplicate rows
- removing rows with retweets
- combining the 3 dataframes into a single dataframe

After cleaning the data and merging the 3 dataframes, I went ahead to analyze and visualize the combined dataframe. The final data contains 1350 tweets. The combined dataframe was stored as a new pandas dataframe called twitter_archive_master.csv