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. The steps carried out in the data wrangling process are briefly discussed below.

1. Gathering Data

In the data gathering process, I was able to gather 3 files for the analysis from WeRateDogs. These files are as follows

- **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.txt that contains the JSON data for each tweet's retweet counts and favorite counts

2. Assessing Data

In the assessment step, the dataframes gathered were assessed both visually and programmatically. The dataframes were assessed for quality issues and tidiness issues.
The following quality and tidiness issues were observed:

Quality Issues

1. The "time_stamp" column in the we_rate_dogs table should be of a Date data type.
2. There are invalid ratings in the rating_numerator and rating_denominator. For example 960/00 is not a valid rating.
3. The "jpg_url" column in the image_prediction table has duplicates.
4. The "name" column in the we_rate_dogs table has some invalid names.
5. The HTML tags in the "source" column in the we_rate_dogs table and twitter_data table obstructs easy reading, and should therefore be removed.
6. Columns like the retweets information column, img_num column in image predictions table are not needed and needs to be dropped.
7. Remove enteries in the image predictions table that have p1_dog, p2_dog, & p3_dog values set to false because these are not dogs of any kind.
8. Correct rating_numerators with decimals

Tidiness Issues

i. The columns ["doggo", "floofer", "pupper", "puppo"] in the we_rate_dogs table need to be combined to one column named dog category.
ii. The we_rate_dogs, image_prediction, and twitter_data tables need to be combined.

3. 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, I went ahead to analyze and visualize the new dataframe. The final dataset contains 1131 tweets and was stored as a new pandas dataframe called twitter_archive_master.csv