Wrangle Report Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This report briefly describes my wrangling efforts.

Project details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

The data for this project were gathered from three different datasets, which are;

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded.
- The tweet image predictions were hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- Twitter API & JSON files: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's ID in JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweets_json.txt. This file was read line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing data

I assessed the data as follows:

- Visually, I used two tools. I imported each of the three datasets in Jupyter Notebook and check the csv file in excel.
- Programmatically, by using different methods e.g. info, value_counts, sample, duplicated etc

Cleaning data

Quality Issues

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float.
- retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- The rating_numerator and rating_denominator columns have incorrect values.
- Missing values in dataset
- Name column have incomplete names e.g 'Mo', 'a', 'an'
- We only want original ratings (no retweets) that have images.
- We may want to change this columns type (in_reply_to_status_id, in_reply_to_user_id,
- retweeted_status_id, retweeted_status_user_id and tweet_id) to string because We don't want any operations on them.
- Drop columns not needed in our analysis
- Some tweet_ids have the same jpg_url, so we will drop duplicates

Tidiness

- We need to Join 'tweet_info' and 'image_predictions' to 'twitter_archive'
- We don't need all information in image_predictions
- We merge all the dataframes into one single dataframes

Cleaning

- We will clean our dataset, and have it tidy up for visualization.

Conclusion

Data wrangling is a core skill that whoever utilizes and handles data should be very familiar with. In this project, I have used Python programming language and its corresponding libraries to wrangle and analyze datasets of different file extensions. The libraries associated with python have made visualization and code documentation very simple and easy. This project on data wrangling have shown that python is a very powerful language. I see the need to proceed into more powerful usage of python for machine and deep learning.