Wrangle Report

Background

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog[1]. In this project, data gathered from three different sources, assessed, cleaned, and analyzed. This report will demonstrate data wrangling efforts in the three main tasks (Gathering, Assessing, and Cleaning).

Gathering Data

Datasets were obtained from three different sources:

1. CSV File of The WeRateDogs Twitter archive. Which was easily read in the Jupyter notebook using pandas.
2. A file of Tweet Image Predictions hosted on Udacity's server. Data accessed using the Requests library.
3. Additional data on Retweets, Count, and Favorites ("like") was obtained using Python's Tweepy Library. Twitter API keys, secrets, and tokens were used to access Twitter API. Then tweets were saved in a jason format. The library Jason was used then to read Jason file and to save it as a dataframe.

Assessing Data

After gathering all datasets, data were assessed as follows:

- Visually: by printing data frames in Jupyter notebook.
- Programmatically: using different methods such as (.info, .describe, value_counts, etc)

Data issues encountered were then assigned as either quality or tidiness issues.

Cleaning Data

Following identifying data issues, is actually handling these issues. This phase included viewing each data set and encountering all quality and tidiness problems.

One of the main cleaning tasks were removing retweets and keeping only original tweets for both archived tweets and tweets obtained using the API.

Moreover, one of the quality issues which havebeen fixed is that some denominators have values greater than 10. Another major step, which was done in the cleaning phase, is replacing the 9 columns of image predictions and confidence levels with only two main columns Dog Type and Confidence Level.

[1] https://en.wikipedia.org/wiki/WeRateDogs

Cleaning also included other regular steps such as: removing duplicates (Image URL), and Changing data types (Tweet ID).
Final step was to merge all data frames into one main dataset.

Packages and Libraries used

- pandas
- numpy
- requests
- tweepy
- json
- time