Wrangle Report

This project was focused on the twitter account WeRateDogs and its associated data. The brief specified that it was to be broken into three distinct stages; Gathering, Assessing and Cleaning. The 3 sources were required, the first being a csv file called twitter_enhanced which held tweet ids, sources, tweet content and the results of some basic text mining. This was separated into dog ratings and 'stages' for the tweets that mentioned it. There were some important omissions, however. The favourite count and retweet count for the individual tweets were missing and had to be gathered through the use of Twitter's API. This was the second source. Finally, the third dataset was a tsv file containing dog breed predictions based off a neural network that took the tweet images as input. This file has to be scraped directly from Udacity's website.

Gathering

- The twitter_enhanced csv file was downloaded manually from Udacity's website and loaded into Jupyter notebooks as a Dataframe with pd.read_csv()
- The missing favourite and retweet information was gathered as a json file using from Twitter's API using the python library Tweepy along with the tweet_id column values from the twitter_enhanced dataframe. Some initial cleaning was done at this point to remove retweet and reply tweet_id's from the dataframe to reduce collection time and extra cleaning after it finished querying the API.
- The image_predictions tsv file was scraped from Udacity's website using the Python requests library, the file was then loaded as a Dataframe using pd.read_csv(). The only change that had to be made was the addition of a separator value '\t', as the initial function call loaded all of the data into a single column.

Assessment

The brief required that at least 8 quality issues and 2 tidiness issues be recognised and dealt with. After programmatically and visually assessing the data, there appeared to be many more issues than the minimum required. Incorrect datatypes, missing values, uncleaned text and unrequired columns were prominent.

Cleaning

- Dropping columns – quite trivial, this was usually done first to reduce the visual clutter when completing the other cleaning steps
- Cleaning text – this was by far the hardest step. My knowledge of strings in Python is lacking and I found the use of regex functions daunting. I eventually got it to work in a disproportionately large time compared to the rest of the cleaning
- Merging columns – This was done to the dog stages as they had separate columns, I originally tried appending them separately, but I never worked out the correct way. The only solution I found was to mine the original text and place the values into a single column. Before dropping the previous columns.

- Fixing incorrect datatypes – this was largely trivial. The only exception was in loading the API json data. Pandas read in the data column as epoch time and an integer datatype. After some research, I realised this and fixed it with an extra parameter specifying the time type.
- Merging datasets - I made the decision to keep the image prediction file and the twitter data separate as they were largely different sets of data. The way I structured the image_prediction dataset was in a relation table layout, that had tweet_id as a foreign key. This was the result of collapsing the dog prediction columns from six into three. I saved it as image_predictions_master.csv. The favourite and retweet columns I merged with the twitter enhanced dataframe on the tweet_id column and saved it as twitter_archive_master.csv