Data Wrangling Report

Gathering data

1. **Twitter archive file:** downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv
2. **The tweet image predictions**, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. **Twitter API & JSON:** Each tweet's retweet count and favourite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing the data

I used pandas .describe() , .value_counts() .sample() and .info() mainly to assess the data. I didn't quite know what 'one' quality or tidiness issue was so I had to make some executive decisions. The issues I found with the data were as follows:

Quality

Completeness, validity, accuracy, consistency (content issues)

twitter archive

1. Deleted columns that won't be used for analysis also that have so many null values('source','in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp', 'expanded URLs')
2. Erroneous datatypes melted them into single column (doggo, floofer, pupper and puppo columns)
3. Separated timestamp into day - month - year (3 columns)
4. Corrected numerators with decimals
5. Corrected denominators other than 10:

   Programmatically (Tweets with denominator not equal to 10 are usually multiple dogs).

image prediction

6. Dropped duplicated URL of jpg
7. Created 1 column for image prediction and 1 column for confidence level
8. Deleted columns that won't be used for analysis

Tidiness

1. Changed tweet_id to type int64 in order to merge with the other 2 tables
2. All tables should be part of one dataset