Gathering the data:

This project began with me gathering the data via URL provided by Udacity. With the URL, I downloaded the image prediction .tsv file. I would need this file because it has the results of the neural network with the proper dog predictions.

Udacity also provided twitter_archive.csv this file has twitter_id which will allow me to capture more data via programmatically and the module tweepy.

After authorizing myself on twitter I have full access to get the status of tweet data .json file, and export for later use. I would have to loop over the twitter ids and export and append them to a list then export that list as a .txt file.

I then created a function called get_tweet. This function is responsible for taking in a list of 3 keywords, and the data JSON file path. after input of the selected keywords, one would get back that data JSON file and tweet_dict which is a dictionary with keywords with its relative JSON file.

Assessing the data:

After obtaining the tweet_dict I was ready to do my visual assessments of all datasets in the data frames. I started off by doing visual assessment first this would help me identify which columns have incorrectly duplicated, or misspelled names. In the course, the instructor talked about assessing programmatically as well. I used the pandas .info module method to help me quickly identify whats corrupt or missing data.

When I completed my assessments I wrote down the various quality, and tidiness issues I needed to take care of. Then it was time to clean my data.

Cleaning the data:

I cleaned the data by combining all three data frames and dropping any redundant columns that are unneeded, and any columns with "Not a Number " values. I changed ratings to be in 1 column instead of two, and drop columns that don't have image prediction belonging to it.

Storing the data:

After completing the cleaning I stored the data frame on disk called twitter_archive_master.csv