At the end of the module on data wrangling taught in the Udacity Data Analyst Nanodegree program, we were given real life data to wrangle and analyze to produce meaningful insights. This would involve gathering data from multiple sources, cleaning the data, and visualizing. All of this enables you to create interesting and trustworthy analyses and visualizations.

The data to be used are to be gathered from three different sources, using different gathering methods.

Twitter Archive CSV File

The first is the tweet archive of the WeRateDogs Twitter account. This twitter account rates people's dogs with a humorous comment about the dog. The ratings had a peculiar rating system. The denominator was almost always 10 and the numerators were usually greater than 10. So you had 14/10, 11/ 10 and so on.

The twitter archive to be used is gotten from Udacity as WeRateDogs had downloaded their twitter archive and sent it to Udacity for the purpose of this project.

All that was needed to do was to read the csv file using the pandas read function. This tweet archive had two notable column omissions. The retweet count and favorite count were absent which led us to the second source.

Twitter API

In order to get the retweet and favorite count, we need to query Twitter’s API Using the tweet IDs in the WeRateDogs Twitter archive. We then use the tweepy library to store each tweet's entire set of JSON data in a text file.

After which the text file is read line by line into a pandas DataFrame.

I was unable to access to access the data using a Twitter developer account, so I just went ahead to read the text file(tweet_json.txt) already provided.

Image Prediction TSV File

The third dataset is an image prediction file. It contains dog breed predictions for all the dogs present in every tweet. This was created by Udacity, using a neural network. This file is a tsv file and is downloaded using the requests library and the ‘\t’ separator.