Introduction

This project was a part of the data wrangling core curriculum of the Data Analyst Nano-degree program of Udacity. Real-world data rarely comes clean. Using Python and its libraries will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog.

Libraries

The following packages (libraries) need to be installed. We can install these packages via conda or pip:

- pandas
- NumPy
- requests
- tweepy
- json

Project Details

For this project, wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

The requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.

The tasks for this project were:

- Data wrangling, which consisted of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting on my data analyses and visualizations (act_report.pdf)

The Data

WeRateDogs provided their Twitter archive to use in this project – twitter_archived_enhanced.csv – Which contains columns: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo). These columns need to be assessed and cleaned.

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. Then read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

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 be downloaded programmatically using the Requests library.