Introduction

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user dog_rates. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering Data

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1. The WeRateDogs Twitter archive
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 should be downloaded programmatically using the **Requests** library and the following URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**
3. Each tweet's retweet count and favorite ("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 DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing and Cleaning Data

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

Quality issues:

df_1:

- timestamp and retweeted_status_timestamp are of type 'object'
- remove columns that are not needed for analysis
- delete retweets

df_image_clean:

- missing data (2075 rows instead of 2356)
- remove duplicated rows of jpg_url

- p1,p2,p3 inconsistent capitalication (sometimes first letter is capital)

df_tweet_json_clean:

- missing data (2354 rows instead of 2356)
- convert tweet_id into int64

Tidiness issues:

- Three data frames should be combined in one data frame since they all describe one tweet

df_1:

- one variable in four columns(doggo,floofer, pupper, and puppo)

Storing, Analyzing, and Visualizing Data for this Project

Store the clean DataFrame(s) in a CSV file with the main one named `twitter_archive_master.csv`.