Data Wrangling Report:

Gathering Data:

About the Dataset(s) the dataset I'll be wrangling is the tweet archive of Twitter user @dog_rates (https://twitter.com/dog_rates), also known as WeRateDogs. This archive/dataset consists of 2356 basic tweet data from November, 2015 to August, 2017. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Based on the images in the above dataset (i.e. WeRateDogs Twitter archive), another dataset is created which consists of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). Though no wrangling will be done directly on this image predictions dataset, it will definitely provide some additional data for our main tweet archive dataset.

Gathering Twitter Archive CSV File:

Using the link provided by Udacity, I downloaded the WeRateDogs Twitter archive manually as twitter-archive-enhanced.csv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv) file and imported this file into a dataframe (twitter1).

Gathering Tweet Image Prediction:

I downloaded the tweet image predictions file hosted on Udacity's servers programmatically using Python's Requests library and saved it locally to image_predictions.tsv file. Then, I imported this file into a Python Pandas dataframe (image).

Gathering Data from Twitter API:

Using the tweet IDs in the Twitter archive, I accessed the entire data for every tweet from Twitter API and stored every tweet's entire set of JSON data in a file called tweet_json.txt file. Created a dataframe twitter2 from this JSON including only tweet_id, retweet_count, favorite_count data.

Assessing Data:

Via assessing the data, I managed to find out some issues in the data which are listed as follows:

Listed Quality Issues:¶
Dataframe twitter1:
1.Data contains retweets (ie. rows where retweeted_status_id and retweeted_status_user_id have a number instead of NaN)
2.tweet_id is an integer
3.timestamp and retweeted_status_timestamp are currently of type 'object'
4.name has values that are the string "None" instead of NaN
5.Some names are inaccurate such as "a", "an", "the", "very", "by", etc. Looking visually in Excel, I was able to find more names that are inaccurate including "actually", "quite", "unacceptable", "mad", "not" and "old. It seems like the method used to extract the names was using the word the followed "This is..." and "Here is..." which leads to some inaccuracies.
6.Found an instance of a name being "O" instead of "O'Malley"

7.doggo, floofer, pupper, and puppo have values that are the string "None" instead of NaN
8.Upon visual inspection in Excel, there are ratings that are incorrect. I ordered the ratings from low to high and looked at the extremes only for incorrect ratings therefore there are likely more than I missed and will be difficult to find them all programmatically. Examples where things may have gone wrong is the use of decimals, or when two instances of numbers separated by a slash are present in 1 text and I assume the first was chosen. Also, there are ratings with decimals such as 13.5/10, 9.5/10 have been incorrectly extracted as 5/10 (in addition to other numbers with decimals such as 11.26 and 11.27). There are instances of 1/2 and 50/50 which are not ratings such signifying "half" which have been considered as ratings. Finally, use of 4/20 and 24/7 has been confused as ratings.
9.There are many columns in this dataframe making it hard to read, and some will not be needed for analysis

Dataframe twitter2:
There are 11 missing tweets compared to the twitter1 datagrame (Might have been deleted)

images Dataframe
1.There are 2356 tweets in the twitter1 dataframe and 2075 rows in the images dataframe. This could mean that there is missing data, or that not all 2356 of the tweets had pictures.
2.tweet_id is an integer
3.p1, p2, and p3 contain underscores instead of spaces in the labels

Listed Tidiness Issues:
Dataframe twitter1:
1 variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

Dataframe twitter2:
twitter2 data should be combined with the twitter1 data since they are information about the same tweet

Images Dataframe:
images data could be combined with the twitter1 data as well since it is all information about 1 tweet

Cleaning Data:

I created copies of the dataframes for cleaning i.e. twitter1_clean = twitter1.copy(),twitter2_clean = twitter2.copy(), and images_clean = images.copy().Then I merged the twitter1, twitter2, and images dataframes on 'tweet_id and named it twitter.Then I managed to clean the issues listed above in the assessing data section. For each quality/tidiness issue, I performed the programmatic data cleaning process in 3 stages - Define, Code & Test.

Storing and Visualization:

After completing the cleaning, I stored the dataframe as twitter_archive_master.csv and managed to showcase some visualization for e.g. favourite vs retweet count.