Data Wrangling Report

1. Gathering Data

The dataset I'll be wrangling is the tweet archive of Twitter user @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.

Gather Twitter archive CSV file

Using the link provided by Udacity, I downloaded the WeRateDogs Twitter archive manually

Gather tweet image predictions

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 (img).

Gather 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 status_df from this JSON including only tweet_id, retweet_count, favorite_count .

2. Assessing Data

Quality

twitter_data

- tweet_id is int
- timestamp should be datetime not str
- in_reply_to_status_id should be str no float
- in_reply_to_user_id should be str no float
- retweeted_status_id should be str no float
- retweeted_status_timestamp should be datetime not str
- rating_denominator should be float
- in_reply_to_status_id: 2278 missing values
- in_reply_to_user_id: 2278 missing values
- retweeted_status_id :   2175 missing values
- retweeted_status_user_id : 2175 missing values
- retweeted_status_timestamp : 2175   missing values
- expanded_urls had a few misiing value
- name columns had some not accurate names like a ,an,the,None
- doggo, puppo, pupper and floofer have many values set as 'None'

Img

- p1, p2, p3 inconsistent, it had some captial words and other small
- tweet_id is int

tweet_api_data

- id is int

Tidiness

- doggo, puppo, pupper and floofer refer to the type of dog and should be in one type column
- twitter_data , img and tweet_api_data should be merged into one dataframe

3. Cleaning Data

For each quality/tidiness issue, I performed the programmatic data cleaning process in 3 stages - Define, Code & Test. During the cleaning process.

1- tweet_id , in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id are int type and it should be str.

2- timestamp ,retweeted_status_timestamp are object type and it should be datetime.

3- rating_denominator should be float because in future dog ratings could have a number with a decimal in the denominator

4- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp had alot of missing value so we should drop it

5- expanded_urls had a few misiing value, as tweet_id is the last part of the tweet URL after "status/" so we will make them

6- name columns had some not accurate names like a ,an,the,None so we should replace it.

7- tweet_id in img is int type and it should be str

8- p1, p2, p3 inconsistent, it had some captial words and other small

9- Id in tweet_api_data is int type and it should be str

10- doggo, puppo, pupper and floofer refer to the type of dog and should be in one type column

11- twitter_data and img,tweet_api_data should be one dataframe so we will merge them

Storing Data

After the completion of the cleaning process, I stored the archive_clean DataFrame in

Twitter data.csv file.