WeRateDogs Twitter:

Wrangling dogs and tweets

A step by step guide to the process used in the report

I was provided with a copy of the twitter.json, twitter-archive-enhanced.csv, and image\_predicitions.tsv for analysis. These were loaded onto my pc for analysis, primary coding was done inside PyCharm Community, then transferred directly to the Jupyter notebook.

### Gather process:

The file that was provided twitter\_archive\_enhanced.csv was directly downloaded and placed into the website.

An image\_predictions .tsv file was obtained using the requests library and taken from the cloudfront host.

Finally a tweet\_json.txt file was gathered using code in PyCharm and the file was then uploaded to the Jupyter workspace. This text was created using research of my own and not the script provided.

### Assessment:

Starting with visual assessment of the JSON file, we identify key structures we want to extract from this file.

* “id”
* “retweet\_count”
* “favorite\_count”

These were then extracted from the JSON data and put into a dataframe, I had considered putting other columns but realized they were not useful for any analysis as we are pulling data from a single source.

During this time I noticed the following issues in quality and tidiness:

#### Quality

* Large number of NaN's in multiple columns
  + 'in\_reply\_to\_status\_id',
  + 'in\_reply\_to\_user\_id',
  + Retweet\_status\_id
  + Retweet\_status\_user\_id
  + Retweet\_status\_timestamp
  + ‘expanded\_urls’
* timestamps are not in datetime data type
* duplicate dog names may or may not be same dog
* Mistyped data types for four categoricals ‘doggo’, etc.
* Maximum rating for a dog is 1776?! Unless this is George Washington’s dog, it probably isn’t right
* All rating denonimators should be 10, but we have at least one greater than 10
* Tweet\_id in df1 and df2 should be string not int
* Img\_num should be string not int
* Df2 contains image guesses that are not dogs.
* Retweet\_count and favorite\_count should be int not object

#### Tidiness

* No data for categorical options doggo, floofer, pupper, puppo
  + These 4 categories should be a single column
  + df1 contains multiple observational units

### Cleaning process:

First I created a copy of df1 called df1\_c . this will be the copy that all cleaning operations are worked on.

I went through our list and located 3 columns to be removed. All 3 are retweet status related and since I am not considering retweets in this analysis, I prepared to remove them. In order to do that I first had to locate all the retweets, I did this by first converting NaNs to zeros and then locating and dropping. I used a ‘purge’ dataframe (df1\_p) so that if I should find that I need this information for some other operation later, I could easily restore it.

After that I went about correcting datatypes, I started with timestamp, then altered all columns that need to be strings and aren’t into the appropriate datatype. I also included df2\_c’s id column for alteration.

To alter category to “none”, “doggo” etc… I chose to first verify that no one dog has more than one category assigned to it.

### External References:

<https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/twitter-data-in-python/>

<http://docs.tweepy.org/en/latest/code_snippet.html>

<https://medium.com/@I_am_milica/the-beginners-guide-to-downloading-twitter-data-using-tweepy-4ec981eaba77>

<https://stackoverflow.com/questions/18869688/twitter-api-check-if-a-tweet-is-a-retweet>

<https://realpython.com/pandas-settingwithcopywarning/>