Project Detail

This project is a data wrangling of The WeRateDogs Twitter archive. There are three steps in wrangling of this project:

- **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.

  3. The tweet json, each tweet's retweet count and favorite ("like") count at minimum.

- **Assessing data**: After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. There are 11 quality issues and 2 tidiness issues in this project:

  - **Quality issues:**
    - twitter archive table:

- Remove retweeted rows, therefore we just keep all rows that have retweeted_status_id is null and drop those retweeted columns.

- timestamp columns are object instead of datetime in archive table.

- name has values that are the string "None" instead of NaN.

- Besides, looking programmatically, 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. So I'll replace all inaccurate names with NaNs.

- Moreover, I saw that has a name being "O" instead of "O'Malley"

- rating_numerator in archive table should be floats, not integers.

- Tweet_id: 786709082849828864 has an incorrectly extracted rating (the value should be 9.75 but 75 was recorded).

- Inconsistency related to "expanded_urls" column in the archive table.

○ image predictions table:

- Inconsistent data: lowercase and uppercase names for p1, p2 and p3 columns in image prediction table.

○ tweet json table:

- Retweet_count and favorite_count in json table should be integers, not floats.

- For easier to read, it should drop undesired columns such as: 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'img_num'

● **Tidiness issues:**

- There are four different columns of dog stages, it should be merged into one column.

- archive, json and images data should be combined together since they are information about the same tweet.

❖ **Cleaning data**:

Clean each of the issues that are documented while assessing by viewing the info columns or checking the value counts of columns, ...