Reporting: wragle_report

Data Gathering:

This project required gathering three data sets. The method used to gather each data was different and are as follows.

- Twitter archive file: This can be downloaded manually or programmatically with the use of the Request library
- The tweet image predictions: This can only be downloaded programmatically using the Request library because the file `image_predictions.tsv` is hosted on Udacity's servers and cannot be accessed manually.
- Tweets: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data found to be interesting are scraped. This is done by:
  - Extracting the tweet IDs in the WeRateDogs Twitter archive and store in another file (tweet_id.txt)
  - Quering the Twitter API for each tweet's JSON data using Python's Tweepy library and store the data in another file (tweet_json.txt)

Data Quality Issues

In the `archive` table

- Change the datatype for some of the columns e.g timestamp
- A lot of missing data in the features
- Missing values represented as `None`
- Expanded_url containing more than one url

In the `image` table

- Lowercase for P1, P2, and P3 sometimes
- Text column not properly formatted

In the `tweet` table

- Extract the date from Created_at column
- Rename the Created_at column as Timestamp to bridge uniformity

Data Tidiness

- P1, P2, and P3 should be formatted properly in the `image` table
- Remove html tags form the source column in the `archive` table
- Tweet_id in `archive` table duplicated in `image` and `tweet` tables

A new data set named 'twitter_archive_master' was produced by merging the three data sets named above, on tweet_id. While uniformity of the column names is crucial for readability, it is also important to enable merging of the datasets. Note that the source link in the `archive` table was dimmed neccessary to be extracted from the html tag so as make it more humnan readable and

loadable in the browser. Although, not all the data quality and tideness issues were addressed (e.g melting the stages of the dog: `doggo` , `puppo` , `floofer` and `pupper` should be in a single column), most all the important cleanings were done.

Details

- There are a lot of missing data in the `archive` table such that `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_id` , `retweeted_status_user_id` and `retweeted_status_timestamp` columns which contain little or no meaningful data should be dropped.
- As for the need for change in the datatype, the following columns were changed
  1. `archive` : Timestamp is a datetime and not object
  2. `archive` : Tweet_id is an object not an integer
  3. `image` : P2_dog is a boolean and not integer
  4. `tweet` : Created_at is a datetime and not integer
- For the column name formating, the column formally labelled as `created_at` was changed to timestamp in `tweet` table
- Formating `p1, p2, and p3` in the image table, dash separating the words was replaced `"-"` with space (`" "`).
- In the `archive` table, `timestamp` and `source` were properly formatted and rewritten in such a way that is much more readable and tidy. These include:

  1. Removing the html tags from the `source` column
  2. Making the timestamp to contain year, month and day only
  3. Choosing only the expanded url that follows the normal pattern
- After merging the table to create the master table, two columns, timestamp_x and timestamp_y, were found to be the same and one was dropped.
- Lastly, two additional columns were engineered as it was dimmed fit that they will be reqiured in answering the research questions

Type *Markdown* and LaTeX: α^2