Wrangling Report

Data wrangling is the process of gathering data, assessing its quality and structure, and cleaning it before do anything like analysis and visualizations.

Gathering:

I've gathered each of the three pieces of data in a Jupyter Notebook titled `wrangle_act.ipynb` into a separate pandas DataFrame:

1. The WeRateDogs Twitter archive through file called `twitter_archive_enhanced.csv` as `df`.
2. The tweet image predictions file called: `image_predictions.tsv` as `df2`.
3. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file as `df3`.

Assessment:

Assess data for:

1- Quality: issues with content.

`df` *table*

- `in_reply_to_status_id` and `in_reply_to_user_id` are int not float.
- `timestamp` is a datetime not a string.
- Replace name for rows that have a given name of 'a' with 'an'.
- set all `rating_denominator` value by 10.
- For `tweet_id` 666411507551481857 `rating_numerator` is 12 instead of 2.
- Capitalize the first character `name` column.
- Nulls represented as None in `name` column.
- Drop unneeded columns for retweets.

`df2` *table*

- For `tweet_id` 667866724293877760 p1_dog is Ture instead of False.
- Add `breed` column.
- Drop unneeded columns.

2- Tidiness: issues with structure that prevent easy analysis.
    - One variable in four columns in `df` table (`dog_stage`) includes (doggo, floofer, pupper, puppo).
    - Merge `df` and `df3`, joining on `tweet_id`, as `df3` is not representing an observational unit.

Types of assessment:

- Visual assessment: scrolling through the data in a Jupyter Notebook.
- Programmatic assessment: using code to view specific portions and summaries of the data.

Clean:

I've started the cleaning stage with creating three new data frames, to apply the cleaning steps, as `df_clean`, `df2_clean` and `df3_clean` alternative of `df`, `df2` and `df3`

Types of cleaning:

1- Manual.
2- Programmatic.

The programmatic data cleaning process:

1- Define: convert our assessments into defined cleaning tasks.
2- Code: convert those definitions to code and run that code.
3- Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

I've applied those three steps of process on each quality issue and tidiness issue.

Store:

I've stored the two data frames after cleaning as `twitter_archive_master.csv` and `twitter_archive_master2.csv`.

That was a briefly report for headlines steps had been token through my wrangling file `wrangle_act.ipynb` where all of those steps clarified in details.