Reporting: wrangle_report

Gathering data

```
The wrangling began with Gathering the dataset to be used for the analysis.
These datasets were gotten in different ways and they are:
```

- `twitter-archive-enhanced` which was gotten through the `pd.read_csv` method which had over 2300 values
- `image-predictions` which was gotten using the requests library, converted from an html file
- `df_api` I developed some code to create an API object that I used to gather Twitter data. After querying each tweet ID, I wrote its JSON data to a tweet_json.txt file with each tweet's JSON data on its own line. I then read this file, line by line, to create a pandas DataFrame.

Assessing the data

This was done in two ways : **Visually** and **Programmatically**. With these methods , I was able to identify several quality and Tidiness issue in the Dataframe. I was able to assess the dataframes visually using methods like `.head()` and `.sample()`. Programmatically, I used methods like `.info()`, `.duplicated()` and deveral more to be able to detect these issues. And with this, I detected the following issues:

QUALITY

1. `twitter-archive-enhanced` The expanded_url column has several missing values in the dataset.
2. `twitter-archive-enhanced` The expanded_url column has duplicated values in the dataset.
3. `twitter-archive-enhanced` There are retweeted tweets on the dataset.
4. `twitter-archive-enhanced` The replied tweets are causing incorrect values on the dataset.
5. `twitter-archive-enhanced` The rating_numerators and rating_denominators have invalid values.
6. `twitter-archive-enhanced` Source values don't have the best quality.
7. `twitter-archive-enhanced` There are names that have one character.
8. `image-predictions` Missing values in the dataset compared to the Twitter enhanced data.

Tidiness

1. `twitter-archive-enhanced` timestamp is not in the right format.
2. `twitter-archive-enhanced` doggo, floofer, pupper and puppo are taking a lot of space.
3. `image-predictions` Incomplete without the archive enhanced data.
4. `image-predictions` p1, p2, p3 are taking a lot of space.
5. `df_api` incomplete withou the archive enhance data.

Cleaning the data

After assessing the data, It's time to deal with the issues being assessed. The problems were solve each respectively:

QUALITY

`twitter-archive-enhanced`

1. Remove all rows that have null expanded url values
2. Clean expanded url duplicates that have RT at the beginning of the text.
3. Drop replied columns in the dataset using `drop` method.
4. Remove all the retweeted columns in the dataset and remove all retweets related columns.
5. Change invalid numerator values to the correct value and denominators to 10.
6. Remove the html tags from the string and leave the content.
7. Correct the invalid names to None.

`image-predictions`

1. Add the best confodence level of the predictions for a dog on one column and delete the remaining columns

TIDINESS

`twitter-archive-enhanced`

1. Change the timestamp format to datetime using `to_datetime`.
2. Melt the puppo, doggo, floofer and pupper columns into one column called dog_stage and drop the variable column and the dupliactes of the dog_stage column
3. Join the enhanced_archive dataframe with the api dataframe using `merge` with the tweet_id as the common point

`image-predictions`

1. Merge the image predictions dataframe to the master datframe with the `merge`.
2. Delete the irrelevant columns.

After this, A few changes were made to the master dataframe especially to the formats of some columns and a new column of the rating ratio was created. This completes the wrangling of the Dataset.

In [ ]: