Wrangling Report, *28 / 07 / 2022*:

- This Project uses 3 datasets related to the the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with and makes funny comments about their dogs.

Three data sources were used:

1. Enhanced Twitter Archive (of WeRateDogs), with 2356 observations.
2. Image Predictions File -generated after running the images of all the dogs through a neural network.
3. Additional Data via the Twitter API.

Importing Libraries.

- I imported various useful libraries, both at the beginning and in the course of wrangling. These include: numpy, pandas, requests, json, io, datetime, os, math, pathlib and functools.
- Later I imported matplotlib.pyplot from matplotlib and seaborn

Data Gathering

- I gathered data from three sources:

1. The Twitter archive dataset downloaded manually from Udacity classroom.
2. Image Predictions file downloaded via the `requests` library from the cloudfront.net url provided in the classroom.
3. I made several attempts to get approval for an authentication key as a twitter developer all to no avail. I therefore used the tweet_json.txt file made available on Udaciy for the additional tweet data dataset.

Data Assesment for Quality and Tidiness:

- I employed both non-programmatic (spreadsheets, pandas) and programmatic directed and non-dorected visual assessment.
- I used pandas methods like dataframe.info(), dataframe.head(), dataframe.sample(), etc.
- I was able to pick out a number of tidiness and quality issues, including: inaccurate dog breeds in the image predictions dataset, the problem of nested columns dataset, irrelevant columns, wrong data types, composite columns, inaccurate numerator_rating

IN DETAIL, THESE ARE THE ISSUES THAT I FOUND:

TIDINESS:

1. The three datasets need to be merged and duplicate columns expecially duplicated id columns and datetime columns dropped.

2. 'floofer', 'doggo', 'puppo', and 'pupper' columns in the Twitter archives dataset 'df_archive' should be merged into one 'dog_stages' column.
3. The display_text_range column of df_img (Image predictions) contains lists which is not suitable for analysis. The upper range should be extracted and set as the value of the column.
4. Nested columns in the Additional information dataset cannot be used directly for analyses. They need to be either normalized or dropped.

QUALITY:

- The WeRateDogs twitter archive `df_archive`

1. Multiple abnormally high rating numerators.
2. Multiple abnormally high rating denominators.
3. The datatype of `timestamp` column should be 'datetime' and not 'object'
4. `tweet_id` column should be a str datatype.
5. Many missing values accross `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id` and `retweeted_status_timestamp` columns.

- The Images Predictions dataset `df_img`

1. `tweet_id` should be a str datatype.
2. Inaccurate dog prediction names.
3. The missing values accross many columns `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, `doggo`, `floofer`, `pupper` and `puppo` columns are falsely 'non-null' because of the 'NANs'

- The Additional data dataset `df_additional`

1. ?

Data Cleaning:

- I first defined the cleaning tasks
- This was followed by coding with adequate, almost step-by-step comments.
- I dropped all nested columns
- I extracted data form some columns e.g twitter client type from 'source' in df_archive
- I corrected some quality errors like the inaccurate rating_numerator column, etc
- I had to download lists of authentic dog breeds to help in cleaning up the image prediction column
- I succesfully combined dog_stages columns after much trying

Data Testing: This was done during and after cleaning to confirm an error-free dataframe using:

- `dataframe.info()`, `dataframe.sample()`, `dataframe.head()`, `dataframe.value_counts()`, etc

RESULTS:

- I ended up with 3 tidy datasets `df_archive`, `df_img`, and `df_additional`
- I merged these datasets using `functools` and got a smaller master dataset with only 1349 observations.
- I proceeded with analyses and visualization using only this dataset

Challenges and Wrangling Problems that I could not Solve:

- The Data wrangling process was time-consuming, required patience and constantly looking up the internet.
- It was extremely tasking and required almost undivided attention
- Even with all of the above, the results were not up to my satisfaction especially because of time constraints.