WRANGLING REPORT

In other to wrangle my data, first I imported I imported the libraries I will be using - *pandas*; for data wrangling, *requests*; used the library to download the *image-prediction.tsv* file and the *json* library; used this library to load the json object from the provided *tweet-json.txt* file.

I then loaded each dataset into a DataFrame and saved in to a variable.

After this, I began assessment on the datasets, beginning with the first dataset; *twitter-archive-enhanced.csv*.

On the *twitter-archive-enhanced.csv*, assessing visually I called the *.head()* method on the DataFrame, I was quick to spot some issues included:

- **the DataFrame appears to have a lot of missing values.**
- **the _expanded*urls* can be dropped as it holds the same values as the ones in the _tweet *id*.**
- **from the *source* columns, the source of the tweet can be extracted and the source of each tweet used to replace the urls.**
- **a lot of missing values in the following columns - _in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status*timestamp*.(can't be clean)**
- **the *doggo, floofer, pupper, puppo* columns can be collapsed into one columm.**

Further performed programmatic assessment on the dataframe and also was able to spot issues that included:

- **the _tweet*id* column should be converted from integer to string.**
- **convert the *timestamp* and _retweeted_status_timestamp_ from object to datetime.**
- **the minimum value for the _rating *denominator* appears to be zero and diving by zero causes an error - ZeroDivisionError.**
- **the *name* column appears to have invalid values and letter/words that is not a name, I performed a random sample on the dataset and examples of inconsitent values can be found in row 801, 924, 2334. These issues concluded my assessment on the *twitter-archive-enhanced.csv* dataset.**

Moving to the next dataset, the _images-prediction.tsv_dataset. Assessing visually I was only able to spot and issue with the dataset:

- **the *p1, p2 and p3* columns have inconsitent form of values - some values in title case and some in lower case. The dataframe appeared to have quality issue.**

Using programmatic assessment, the issue I spotted was the issue of wrong datatype

- **the _tweet*id* column should be converted from integer to string.**

The *image-prediction* dataset is less messy compared to the *twitter-archive-enhanced* dataset. This concluded my assessment on the images-prediction dataset.

And to the last dataset, the _tweet *json.txt* dataset, before assessing this dataset, first I had to read in the *.txt* file, convert to json and stored the json object in a list, then went on to convert it into a dataframe. Carried out visual assessment on the dataframe and the dataset appears to be okay.

When on to perform programmatic assessment on the dataframe and the only issue spotted was

- **the *id* column should be converted from integer to string/object.**

This concluded my wrangling processes.