WRANGLING REPORT

GATHERING:

Data was gathered from three different sources

1. **Source: File on hand (twitter_archived_enhanced.csv)**

   This file was readily available on my machine and was read into a jupyter notebook using pandas function .read_csv().

2. **Source: Downloading file from the internet (image_predictions.tsv)**

   This is a flat file hosted on udacity website. This was downloaded into my machine using the 'requests' module in python.

   It was later read into a jupyter notebook using pandas function .read_csv().

3. **Source: Twitter API**

   Using the access keys gotten from Twitter, I accessed the API to get the jSON files the tweet_ids that were given in the image_predictions.tsv file.

   Using json.dump, each tweets json file was read into a txt file on separate lines.

   Using json.load, they were loaded into an empty list to form a dictionary. The dictionary was then used to get each tweet_id, favorite_count and retweet_count.

   These three columns were used to form a new data set (more_tweet_info)

ASSESSING

All three datasets were assessed visually and programmatically.

For visual assessment, the datasets were displayed in jupyter notebook, and were scanned through.

For programmatic assessment, python functions such as .head(), .tail(), .sample(), .info(), pd.series.value_counts(), .duplicated(), .describe(), amongst other methods / functions were used in assessing the datasets.

After assessment, a list of quality and tidiness issues were identified, to be taken care of in the cleaning section. The list include:

Quality Issues¶

- Numerator on twitter archive dataset has an outlier of 1776
- Denominator has an outlier of 170

- Timestamp and retweeted_status_time_stamp are in string format
- Null Values in Dog stage columns represented as None
- Name field contains 'a' and None which are not names
- Other prediction values are not necessary for analysis since p1 is confirmed as the highest
- The tweets on more_tweet_info dataset without images are not needed
- The tweets on twitter_archive dataset without images are not needed
- Retweeted_status_id and retweeted_status_user_id that are not null indicates retweets or replies which are not needed
- There are some rows with predictions as the images not being dogs

Tidiness Issues¶

- Dog Stages columns should be single column
- twitter_archive dataset contains more than one observational unit. (Ratings and Dog info)

CLEANING:

Cleaning activities were done following the Define, Code and Test steps. A variety of pandas functions were used in the cleaning which includes, but not limited to:

1. .replace function: This was used to replace some values. Like replacing None with np.nan
2. pd.datetime: This was used to convert timestamp format from string to datetime
3. .drop(): Used to drop rows and columns that were not required.
4. np.select: Used to compress the dog_stages columns into one.
5. pd.merge: Used to merge all datasets into a master data set.