Wrangle Report

- **Introduction –**

  Real-world data rarely comes clean. Using Python and its libraries, I will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. I will document wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries)

  The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

- **Gathering and reading all of the three required Dataset –**

  1. Reading WeRateDogs Twitter Archive.
  2. Used Requests library to download Tweet image prediction programmatically.
  3. Downloaded missing tweets data using Twitter API called Tweepy and saved the downloaded json data in tweet_json.txt

- **Assessment of the dataset downloaded and read –**

  1. **Visual Assessment** -

     A. **twitter_archive_df** - In the first look is could see lots of NaN and not much related to our analysis in few of the columns hence they seemed not useful. Also noticed source column contained html tags, and doggo, floofer, etc which are dog stages hence an observation still they were columns heading so it was not tidy.

     B. **image_prediction_df** – Some of the dog prediction were not related to dog breed or even dog hence those had to be handled

C. **extra_tweets_info_df** – retweeted columns looked to just have a single value i.e. False that has to be looked into.

2. **Programmatic Assessment –**

A. **twitter_archive_df –** Using various functions few things that were assessed on this dataset are as follows First, five columns had most of there row set to NaN. Second, timestamp column datatype was string. Third, although dog stages columns seemed to have no null rows but it was not actually so, None was a string hence was being perceived as a non-null row. Many of the assessment was done by me during the cleaning process our even during analysis. Hence, I kept on hopping between the three processes of assessment, cleaning and analysis.

B. **image_prediction_df** – Rows unrelated to dog breed have to be set to None as they are of no use to me.

C. **extra_tweets_info_df** – My visual assessment was correct retweeted column just had one unique value i.e. False

- **<u>Cleaning the Datasets</u> –**

A. **Quality –**

1. Convert timestamp and retweeted_status_timestamp from string to datetime format.

2. Extracting source from text in source column removing html tags in twitter_archive_df_copy

3. Change dog stages from 'None' which is a string to None and handle dog which are assigned more than one stage.

4. Extracting score from the tweets using regex as some score were not extracted properly and score like 24/7, etc were present.

5. Dog names in lowercases are not real hence have to be changed to None as names like a, an, etc were present.

6. Drop columns from twitter_archive_df_copy which will not be useful during analysis.

7. Drop retweeted column from extra_tweets_info_df_copy as during assessment I have seen that the columns has just one value ie. False hence it is not useful.

8. Dropping of retweeted tweets and tweets for which prediction are not available.

B. Tidiness –

1. Using twitter_archive_df_copy.melt() to make a single column for dog stages as column of a dataset can not be an observation.

2. In image_prediction_df_copy we can handle dog breed using two columns ie. breed_type and prediction_confidence.

3. Merging all the three dataframes as they are single observational unit using pd.merge().

- **Saving the merged dataset as twitter_archive_master_df.csv**
- **And now analysis starts…**