PROJECT 2: DATA WRANGLING

The dataset wrangled in this project is the tweet archive of the Twitter user named WeRateDogs whose twitter handle is @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about them (the dogs).

The data wrangling activities of this project involved the three (3) main processes of **Data wrangling** namely; “Data gathering”, “Data assessment” and “Data cleaning”.

Data Gathering:

In this project, data was gathered for three related tables using three different techniques. The first table was gotten from a **.csv** file (named twitter-archive-enhanced.csv) that has already been uploaded to the project workspace. This file was read into a pandas DataFrame named “twitter_archive_enhanced” using the pandas.read_csv() function.
The second table was gotten from a **.tsv** file (named image-predictions.tsv) that was downloaded using the requests library. After the download, this table was read into a pandas DataFrame named “image_predictions” using the pandas.read_csv() function, with **\t** as its separator since it is a **.tsv** file.
For the third table, the tweepy library was used to query data via the twitter API with the corresponding **Json** data of every query being saved into a **.txt** file named “tweet_json.txt”. Three different information on this **.txt** file were then read into a pandas DataFrame named “tweepy_data”. They are; tweet_id, retweet_count and favorite_count.

At the end of the data gathering process, we were provided with three tables (twitter_archive_enhanced, image_predictions and tweepy_data).

Data Assessment:

The data assessment in this project was done both visually and programmatically. The three tables were displayed for visual assessment while the information of all 3 tables was gathered programmatically to enable extensive evaluation.

In the end, thirteen (13) issues were documented with eight (8) being data quality issues while the remaining five (5) issues were data tidiness issues.

Data Cleaning:

In the data cleaning section of this project, the data quality and tidiness issues documented in the data assessment section were all addressed. But before that, a copy was made for each of the three tables involved in the project. twitter_archive_enhanced was copied into **df1**, image_predictions was copied into **df2** while tweepy_data was copied into **df3**. In the end, the cleaned master dataset was saved to a **.csv** file named “twitter_archive_master.csv”.