Purpose of the Report : To satisfy the nanodegree requirements and sharpen my skills in data analysis

Scope : The dataset wrangled (which was later on analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs (an account that rates people's dogs with a humorous comment about the dog.)

The dataset is comprised of three pieces each from different sources:

1. *The twitter archive enhanced dataset*
   This dataset was downloaded manually from a link provided by Udacity's platform and loaded into a data frame for assessment
2. *The tweet image predictions dataset*
   This file (image_predictions.tsv) is present in each tweet according to a neural network. It was hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and then loaded into a data frame as well.
3. *Tweet_Json.txt dataset*
   This file contains each tweet's retweet count and favorite ("like") count amonst additional data. Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API was queried for each tweet's JSON data using Python's Tweepy library and each tweet's entire set of JSON data was stored in a file called tweet_json.txt file. Each tweet's JSON data was written to its own line and its .txt file read line by line into a pandas Data Frame with the columns tweet ID, retweet count, favorite count etc.

After loading the datasets into data frames, A visual and programmatic assessment was done in other to bring out at least two Tidiness issues and eight quality issues. The modules/libraries imported for assessment, likewise cleaning were NumPy, Pandas, Matplotlib, Requests and os.

The following issues were found and documented:

For Quality Issues

- Columns with several empty enteries in the tweet_json dataset were found i.e. *in_reply_to_status_id,in_reply_to_status_id_str,in_reply_to_user_id,in_reply_to_user_id_str ,in_reply_to_screen_name, quoted_status_id_str, quoted_status_id,* and *quoted_status*
- Columns without any data were present i.e. *geo, coordinates, place* and *contributors* columns.
- The *timestamp* and *retweeted_status_timestamp* columns had incorrect datatypes i.e Object instead of datetime
- For the twitter_enhanced-archive dataset, the source column contained extra information in its records (added to the source of the tweet), which wasn't of any use to the analysis but rather added a lot of jitter.
- For the tweet_json dataset, the source column contained extra information in its records as well, which I deemed wasn't of much use.
- Several columns in tweet_json were not of major importance for our visualization and only served as a distraction.

- For the twitter-enhanced column, we only wanted original ratings (no retweets) that had images. Though there were 5000+ tweets in the dataset, not all were dog ratings and some were retweets.
- In the tweet_json data, the column *Lang* contained information about the languages which were abbreviated in a manner not easy to identify which language was being referred to,

As for Tidiness issues:

- For the twitter archive dataset, there were four columns () which were not individually variables but pointed to the same type of observation (dog stage), whereas they should have been values for one column.
- In the image predictions' dataset, the final breed could not easily be assessed directly for visualization hence needed restructuring.
- the datasets twitter_json, image predictions and twitter_enhanced_archive all observed data of the same nature (related data) hence needed to be merged into one table to avoid dealing with several tables.

After the above errors were assessed and documented, each issue was individually cleaned using the Define-Code-Test technique where each issue was summarily highlighted in the **Define phase**, its solution then coded in the **Code phase** and then tested for its accuracy in the **Test phase**. This procedure was repeated for each issue.