WRANGLE REPORT

For the Wrangling and Analyze Data Project, I utilized Python packages which include numpy, pandas, matplotlib, json, requests, and IPython.

GATHERING DATA

The first set of data (twitter_archive_enhanced.csv) was provided by Udacity of which I directly downloaded into my local machine and then read to a pandas dataframe.

The second set of data containing the tweet image prediction was scraped from the URL (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) with the help of the Python requests library. I utilized the open() method to write the scraped data to a file named "image_prediction.tsv" and then read it to a pandas dataframe using the pandas read_csv() method.

For the additional data from the Twitter API, I utilized the "tweet-json.txt" file provided by Udacity because I was not able to secure a twitter developer's account. I read this data line by line and loaded it to a pandas dataframe.

ASSESSING DATA

Visual assessment was done by merely calling the dataframe objects of each dataset, and also with Notepad++.

Programmatic assessment was done using pandas dataframe methods such as info(), duplicated(), describe(), and value_counts(). After the assessment, nine quality issues and two tidiness issues were identified which include:

Quality issues

1. Retweets need to be removed.
2. Timestamp, retweeted_status_timestamp and created_at columns in Enhanced and Tweet tables have object datatype whereas they should be Datetime.
3. The displayed value of first entry in the 'truncated', 'is_quote_status', 'favorited', 'retweeted', 'possibly_sensitive', 'possibly_sensitive_appealable'

columns of the Tweet table is '0.0' whereas it was 'false' when viewed with Notepad++.

4. Null values represented as 'None' in 'doggo', 'floofer', 'pupper', and 'puppo' columns.
5. Incorrect Dog names.
6. Id column datatype in Tweet table is float and not integer.
7. Column name inconsistency ('Id_str' instead of 'Tweet_id' in Tweet table).
8. Invalid rating.
9. Missing data in both Enhanced and Tweet tables.

Tidiness issues

1. 'in_reply_to_status_id', 'in_reply_to_user_id', and 'source' columns in Enhanced table duplicated in Tweet table.
2. 'doggo', 'floofer', 'pupper', and 'puppo' columns should make up a single column which represented all the stages.

CLEANING DATA

The above issues were dealt with programmatically which resulted into clean dataframes. Below is a brief definition of the solutions implemented on the dataframes:

- Issues #1: Subset the enhanced_clean dataframe using the isnull() method to filter the 'retweeted_status_user_id' column.
- Issues #2: Apply the Pandas to_datetime() method on the columns to transform them to Datetime datatype.
- Issues #3: Apply the astype() method to transform the content of the columns to string and then replace the floating point value with 'False'.
- Issues #4: Define a function to replace the values and perform a function call to effect the changes.
- Issues #5: Replace the values of 'None' and other invalid names in the column with NaN.
- Issues #6: Apply the astype() method to transform the content of the column to integer.
- Issues #7: Rename the 'id_str' column to 'tweet_id'.

- Issues #8: Visually inspect the entries with denominator != 10, note down the indexes of entries with invalid rating and drop them.
- Issues #9: Drop columns with more than 50% value missing.
- Issues #10: Drop the duplicated columns in tweet_clean table.
- Issues #11: Concatenate the 4 columns to form 'stage' column and later drop them.

MERGING THE THREE DATAFRAMES

I merged the three dataframes to form a single dataframe using pandas merge() method.

Upon merging, several values were missing in the merged dataframe. I filled the columns with object datatype with “None”, columns with float datatype with “0.0”, and columns with datetime datatype with the median of the respective column.

STORING DATA

I saved the clean dataframe to a file named “twitter_archive_master.csv” using the pandas dataframe to_csv() method.