WRANGLE REPORT

The wrangle effort strictly followed the standard wrangle procedure of, Gather, Assess, and Clean.

GATHERING: For the purpose of this project, 3 different data sets were gathered from three different sources using three different techniques. The first data was a CSV file in hand (Enhanced Twitter archive) provided in the classroom. This was downloaded manually and imported into a pandas data frame. The second file (image prediction) is a TSV file that was downloaded programmatically using the request library in pandas and the OS library was used to create a folder in which the file was stored. The last dataset was gathered by querying the Twitter API using an API object created using keys generated by access to the API and the Twitter API library called tweepy.

ASSESSING: All three datasets were assessed both visually and programmatically. The datasets were checked for quality issues and tidiness issues and below are issues found when checks were carried out.

QUALITY ISSUES

1. Twitter Archive: Retweet and reply data is present in the dataset, only original tweets are required.

2. Twitter Archive: Insignificant columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, name, Doggo, Floofer, Pupper, Puppo) to be removed.

3. Twitter Archive`: Timestamp data type 'object' instead of 'DateTime'.

4. Twitter Archive: Incorrect dog rates.

5. Twitter Archive: Significantly out-of-range ratings, mostly for a group of dogs.

6. Tweet ID datatype is int instead of a string.

7. Image prediction: non-descriptive column headers.

8. Image prediction: Inconsistent format for predicted dog breeds, some were capitalized others were not.

TIDINESS ISSUES

1. Additional twitter data: Favorite count and retweet should be in the Twitter archive table.

2. Twitter Archive: Expanded URLs contain more than one variable.

All issues noticed during the assessment of the data sets were documented to make the cleaning process a lot easier. All were cleaned in the cleaning stage of the data wrangling process.

CLEANING: Before cleaning a copy of all the data sets was made to ensure that the original data was not tampered with.
The retweet data and reply data in the dataset were removed by assigning the data (in_reply_to_status_id and retweeted_status_id) that is not null to two variables and using the drop function in conjunction with the index method to drop them. All insignificant columns were dropped using the drop function and the timestamp datatype was converted using the 'to_datetime' function.

To correct the Dog rates, the original ratings were extracted from the text and saved in column name 'rating' from which the numerator and denominator were re-assigned to the appropriate columns. A couple of ratings did capture correctly as they had more than one data that matched the regex used and as such were corrected manually.

Some ratings were significantly out of range and further investigation by visiting the Twitter page of those tweet IDs revealed that the ratings were mostly for groups of Dogs and to make the data consistent they were dropped.

The tweet IDs were converted to string as a unique identifier of each tweet using the 'astype' method, and the non-descriptive headers were corrected by renaming them.

Finally, on the quality issues, the dog breeds from the image prediction dataset all had their first letters capitalized for consistency.

For the tidiness issue, all data sets were merged into a single dataset and the expanded URL column was cleaned to contain just one variable.