Wrangle Report

The wrangling process occurred involved three pieces of data set. They were loaded into the Notebook as the **Gathering** part of the analysis.

1. The Tweets Archive CSV dataset, as provided by Udacity, which included general information about the tweets and its contents, including tweet id (the primary key in our case), the rating of the dog, and the stage of the dog.
2. The Prediction TSV dataset, as provided by Udacity, which included key information for our analysis on the prediction of the type of dog in each tweeted picture.
3. The Tweet Interaction dataset, although meant to be extracted through Twitter API, the dataset provided by Udacity, which included information such as number of retweets and favorites for each tweet.

I then proceeded with the **Data Assessment** part of the exercise, utilizing:

- Visual assessment through .**head()** method of Pandas.
- Programmatic Assessment through .info(), .value_counts() and other methods of Pandas.

The observed issues in Quality and Tidiness were categorized as following before the **Data Cleaning** process.

1. **Data Quality Issues**
    - **Archive dataframe**
        - Columns with too many missing information that should be deleted, as they are not strongly needed by this analysis, and the dataframe would be cleaner (source,in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id , retweeted_status_timestamp)
        - The null values is also shown as None in various columns including the dog stage and the name column, which bypasses the programmatic assessment.
        - Some rating_denominator data is not 10 (20 in total). A conversion will be necessary.
        - One rating_denominator value is zero, this would need to be removed since it would be problematic for mathematical assessment
    - **Predictions dataframe**
        - jpg_url column might not be necessary, but we will keep it for the time being until there is further clarification.
        - dog names are separated with "_", while this is not a big issue we might consider replacing it with a space instead.
    - **Twitter Interactions dataframe**
        - There are many columns that are unnecessary for the type of analysis we want to do. It's recommended to drop all columns except for (favorite_count, retweet_count, id, id_str)

        - id and id_str should be the same , however and not all of the values are identical, only 64.5% match. While normally one of them should be sufficient, perhaps its better to keep them both for further inspections.
        - id and id_str data types are integers, they should be converted to strings since they are not subject to mathematical assessment.

2. **Data Tidiness Issues**
    - **Archive dataframe**
        - tweet_id is an integer, while this is not really an issue it technically should be a string
        - timestamp is an object (string), should be parced as datetime
        - The doggo, floofer, pupper, poppo columns are values, not column titles. merging them into a new column, let's say "Category" would be useful.
    - **Predictions dataframe**
        - tweet_id is an integer, while this is not really an issue it technically should be a string
    - **Twitter Interactions dataframe**
        - the final id column should be renamed to tweet_id to match the other two columns

The data was cleaned, depending on the type of the issue through methods such as Join, Query, Replace, Drop_Duplicates, and Rename.

Copies of each Dataframe were made before any changes, and the cleaned ones were joined in one Master dataframe which was exported as a CSV file for future uses.