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, downloaded programmatically through Python, 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, I did not have access to the API so I had to use 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**

1. **Data Quality Issues**
    A. **Columns Removal**
        a. ***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 (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.

        b. ***Prediction Dataframe***
            - all will be kept.

        c. ***Tweet Interaction Dataframe***
            - There are many columns that are unneccesary 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)

    B. **Column Rename**
        a. ***Tweet Interaction Dataframe***
            - the final id column should be renamed to tweet_id to unify with the other two dataframes

    C. **Zero rating denominator**
        a. ***Archive Dataframe***

        - One rating_denominator value is zero, this would need to be removed since it would be problematic for mathamatical assessment

D. **Adding new rating column**
    a. ***Archive Dataframe***
        - Since some ratings have different denominator, we might consider creating a new column with calcualted score

E. **Data Replacement**
    a. ***Archive Dataframe***
        - The null values is also shown as None in various columns including the dog stage and the name column, which bypasses the programmatic assessment.

    b. ***Prediction Dataframe***
        - dog names are separated with "_", while this is not a big issue we might consider replacing it with a space instead.

F. **Incorrect Data Types**
    a. Changing IDs from Integer to String:
        - tweet_id in both Archive and Prediction Dataframes to be converted from Integer to String
        - id and id_str in Tweet Interaction Dataframe

    b. Changing timestamp from an object (String) to datetime
        - The timestamp in the Archive need to be changed to datetime format for connection and to enable timeseries analysis

G. **Extracting Meaningful data from within a column**
    a. ***Archive Dataframe***
        - The source column has to be cleaned from an HTML code to extract meaningful information.

H. **Duplicated Data Entry** (from Udacity Feedback)
    a. ***Different dataframe size***:
        - The Image Predicton Dataframe had 2075 entries, compared to 2356 in the Archive Dataframe and 2354 in the Tweet Interaction
        - There are apparently tweets that are duplicated in the dataframe due to retweet. This would need to be filtered out through the final dataframe with rows that do not have images before conducting any further analysis.

2. **Data Tidiness Issues**
    A. **Dog stage variables forming different columns**
        - In the Archive dataframe, the doggo, floofer, pupper, poppo columns are values, not column titles. merging them into a new column, let's say "Category" would be useful.

    B. **Merging the three tables form one observational unit**
        - There are three different tables that should be joined into one for practicality.

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.