Data Wrangling Steps:

Here we have investigated the dataset obtained from Twitter's WeRateDog page.

1. **Gather**
   Here data has been gathered from three different sources.

   1. I have first read the file which was already provided (twitter-archive-enhanced.csv).
   2. Next I have gathered data for images programmatically using requests library. (image_predictions.tsv).
   3. Using tweepy, I have queried Twitter's API to gather more information.
   4. After querying each tweet ID, I have written the data to a JSON file. (tweet_json.txt) . This file is then ready line after line to create a pandas dataframe.

2. **Assess**

   1. First I have made copies of all the three dataframes.
   2. I have checked all the three dataframes for duplicates, data types, shape etc to decide on all the Quality and Tidiness issues that needs to be addressed.
   3. **Tidiness** issues found:
      1. tweets_df should be a part of twitter_archive_df table.
      2. There are multiple columns for dog stages which is redundant. We can include all in just one single column.

   4. **Quality** issues found:

      1. Drop the columns which does not contribute to our analysis .
      2. We are looking to deal with information that do not have missing image information, hence any such missing records.
      3. Convert inapropriate datatype of timestamp column to the correct one. (from object to datetime type) .
      4. Capitalize the first letter of Dog names inorder to maintain consistency.
      5. There are many unusual dog names like " a,an, this,the" etc which seems like a typo. Hence, we will be removing these names.
      6. Changing few column names for better user readability.
      7. Optimizing the source types and removing the extra text.
      8. Removing null values from retweets.

3. **Clean**

Here all the quality and tidiness issues have been addressed.

Tidiness:

I have merged all the three dataframes into one dataframe and all the four dog stages have been melted into one dog_stage column.

Quality issues:

1. All the irrelevant columns and columns have missing image information have been dropped. Also null values from retweets column have been removed.

2. Addressed inappropriate data types.

3. Removed inappropriate dog names and renamed few columns for better user readability.

4. I have also capitalized all the dog names to maintain consistency and removed extra text from source column.