**Overview:**
The Data Wrangling process for the WeRateDog was not a stressfree exercise. I had to gather the required dataset from three different sources. I downloasded the Image_prediction file programmatically, I downloaded the twitter_archive_enhanced.csv file manually, I also tried querrying twitter's API to download the WeRateDog data, but resorted to using the provided JSON file on the classroom after some failure to download the 2356 rows of data set to avoid data quality issues

**Datasets:**
The Image_prediction dataset has 2075 rows and 12 columns, the twitter_archive_enhanced dataset has 2356 rows and 17 columns while the tweet_api file has 2354 rows and three columns(this was as a result of reading only the tweet_id, retweet_count, and favorite_count into a dataframe)

**Assessment:**
During the visual and programmatic assessment, some quality and tidiness issues were noticed about the dataset which includes: incorrect datatypes, non descriptive headers, wrong information on the dog name columns, one variable splitted into four columns, and lots more.

For the reason of this analysis, some of the issues were cleaned so as to have a tidy quality dataset. Before cleaning, I ensured I created a copy of each datasets so as not to alter the form of the original dataset.

**Cleaning:**
I observed that the three datasets had their tweet_id as an int rather than string since there won't be any mathematical operation to be done with the ids. I used pandas .astype() method to convert the datatypes to string passing the argument str into the .astype() method.

Also I noticed that about nine columns from the image_prediction dataset has headers that are non descriptive, hence I changed them to a more relatable header using pandas .rename() method
Using programmatic assessment, I noticed that the names of dog on the dog column differs sometimes from their original name on the text column of the twitter_archive_enhanced table. I heard to view a sample of 10 observations from the text columns so as to find patterns on how to extract the dog names. From my observation, I was able to extract the names of dogs where it exists using regular expressions. After extraction of the names,1541 names were extracted and stored in a new column named dog_name. I noticed there was an improvement as 45 names more were extracted compared to the initial names saved in the name column.

I also changed the timestamp column from object datatype to datetime using pandas to_datetime() method

During visual assessment, I noticed that the dog stages(I.e doggo, floofer, pupper, puppo) were filled with None rather than NaN, hence the need to change it to NaN using numpy np.nan syntax so as to keep the originality of values

I observed that the tweet source column was not well presented as the sources were embedded in HTML tag, hence I formatted the column by extracting the sources using regular expression

Also in the rating_denominators, some values were found to not be in the multiples of 10 such as the numbers 0,2,7,11,15 and16. I decided to replace them with the number 10.

Also the dog rating could be represented in a more better way rather than splitting the numerator and denominator, hence I multiplied by 100 the value when the rating_denominator divides the rating_numerator and stored it in a new column. I dropped the rating_numerator and raating_denominator columns afterwards.

I also dropped some columns that pose as irrelevant to the cause of the project.
I also noticed that the dog stages could be represented in a more better way

I also used pandas wide_to_long method to reshape the 9 columns into three on the image_prediction table, and that made the table looks more simple and understandable.
I ensured that there were no two columns holding same information, afterwards I merged the three datasets before storing as a .csv file