Reporting: Wrangle_report

In the data wrangling process, I worked with three data sets with (2356, 17), (2075, 12), and (2354, 3) rows and columns respectively. I used the following libraries **Pandas, NumPy, Matplotlib.pyplot, Matplotlib, Requests, JSON, and re.**

For the first data in CSV which is **data1**, it has **2354** rows and **12** columns with **5** columns having missing values. To clean the data, I implemented the rule for data tidiness which states that which states:

- Each variable forms a column and
- Each observation forms a row.

This I did by combining all the columns that had all four dog stages into one column called **dog_stages** then I dropped the columns which I had already combined. Next, I dropped the columns with missing values, this is because the number of missing values is too much and has no consequence for our analysis. But before dropping the columns, I, first of all, dropped the rows before the columns because if I dropped the columns first, I would end up removing some records that ought not to be removed. I dropped the rows by first getting the index that had those missing values then dropped the indexes.

Next, I extracted the rating from the text column with re library then I split it into rating numerators and rating denominators then I dropped the old ratings which were wrong.

For **data2**, it has **2075** rows and **12** columns. It contains the tweet image prediction data, the data not only has dog images but also other images. To tidy the data I have to filter all dog images only, these I did with np.select, a NumPy function. After this, I dropped other columns that are not relevant to the analysis.

For **data3**, it has **2354** rows and **3** columns. The data is clean. My next was to merge all three data into one data but before I did that I has to rename the id column to tweet_id, this is because I want the unique columns to have the same naming. After merging the columns on the tweet_id, I did a little data cleaning, first I replaced the underscores in the breed names with space. I charge the dtype of timestamp from object to datetime. Lastly, I dropped some columns which are not relevant to my analysis and made a copy for each data for my analysis.