Wrangle report

Gathering Data:

I gathered the data from databases for the WeRateDogs twitter account and read them using the Pandas Python library in Jupyter workspace. There were three databases to use in this project, all containing the tweet ID.

Assessing Data:

- I made a copy of each database to make sure that all the cleaning efforts are done in a separate data and the original is preserved.
- Then, I noticed that the timestamp column is not easy to read because the date and time are merged together. I separated them into a date column and a time column using the datetime from Pandas.
- After that, the database had three columns for date and time: date column, time column, and the original timestamp column. I dropped the timestamp column using the drop function
- I renamed the id column in df3_clean to tweet_id using the rename function to match the other databases because we will be merging the columns later.
- I removed duplicate images in jpg_url column using duplicated() and drop_duplicates.
- I rounded up the decimals in confidence rate columns to 2 decimals
- I noticed that the names of the predicted dog breed are separated using "_" instead of spaces, so I replaced the _ in the predicted names to spaces.
- I dropped retweets columns
- I dropped replies columns

Tidyness:

- The database had columns for different dog types. Since only one type can be selected, the others just have "None" values. To tidy the database, I merged the columns into one column called "dogs" and changed the values to reflect the selected breed without the "None" values. Then, I dropped the original columns.
- Then, I merged all three databases using "tweet_id" column to have one master database that contains all the columns. Note that I merged the clean databases and not the original.