Abdelrahman Issawi.

Sources
I gathered data from a given CSV, a website, and Twitter's API. I used tweepy to access the API and gather the JSON data for the tweets. I stored the JSON data in a text file, then loaded what I needed into a pandas dataframe.

Process
I audited the data by checking data types, value counts, number of non-null entries, and numeric summaries. Data for a few tweets (5) was unrecoverable. I combined (that is, inner joined) all three tables because each column is a feature of its tweet. I reshaped the dog stages (e.g., floofer, puppo, pupper, doggo) into a single column rather than multiple columns. The few tweets (14) that had multiple stages in the text I put into a new category called multiple. I converted several columns to new data types:

- Tweet IDs, status IDs, and user IDs from integers to strings
- Timestamps to datetime objects
- Dog stages from strings to categories
- Retweets and favorites from floating-point numbers to integers

I manually fixed some dog names that were incorrectly captured from the text (e.g., a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such) by visually inspecting the visible portion of the tweet's text.

I manually fixed some dog names that were incorrectly captured from the text (e.g., a, an, the, just, one, very, quite, not, actually, mad, space, infuriating, all, officially, 0, old, life, unacceptable, my, incredibly, by, his, such) by visually inspecting the visible portion of the tweet's text.

For each fixed issue, I identified the issue, stated my intention, then tested to ensure that I enacted my intention.

Storage
I stored the data in two CSVs, one for the tweet data and one for the image predictions.

Remaining Issues
1. I noticed a strange interaction where a tweet's text was cut off in the dataframe.
2. I decided not to change the numerators or denominators for the dog ratings because they probably were not mistakes. They were humorous deviations. If actual analysis is done, then exclude those values.
3. Some of the variables did not seem necessary or useful (e.g., *in_reply_to_status_id*), but I left them in case someone else could make use of them. The cost of storing the additional data is minimal.