My wrangling efforts were divided into three main tasks, that are gathering data, assessing data and cleaning data. In this report, I will make a brief summary of how I performed each of these tasks:

Gathering:

. Directly downloaded the 'twitter-archive-enhanced.csv' archive provided by udacity and loaded it into a dataframe.
. Downloaded programmatically the image_predictions.tsv provided by Udacity, saved the file as .tsv and then loaded into a dataframe using read_csv with a sep='\t' parameter.
. I could successfully connect to the Twitter API, and retrieve some tweet ids, with their retweets and favorites values. To do this, I had to use .Client(), instead of .API(), because my access wasn't elevated enough. However, since it took to long, and my Jupyter Notebook crashed several times, I decided to use the tweet-json.txt provided by Udacity. Using some code, I could successfully load it into a dataframe.

Assessing:

. I did a visual assessment of the 3 dataframes in the Jupyter Notebook using .head() and sample(25). I also did a visual assessment using Google Sheets, which was useful because the data was easier to inspect visually this way.
. I did programmatic a assessment of the 3 dataframes in the Jupyter Notebook using .value_counts() and .info() in the 3 dataframes. With info() and .value_counts(), I could check if there were any duplicates in the tweet_id columns, but there weren't any. Also, it helped me to get a general understanding of the dataset. Moreover, I also used .value_counts() to check for errors in the columns name, rating_denominator and rating_numerator.
. I also did a programmatic assessment, using .query to check for some strange values in rating_numerator and rating_denominator that could have errors.

Cleaning:

I solved the issues I discovered in my assessment dealing first with the tidiness issues and then with the quality issues.

Tidiness issues:

1. I joined the tables using merge and the column tweet_id. Also, I had to change the data type from integer to string.
2. I merged all the dog stage columns into one using melt.

Quality issues:

1. I changed the values in the new column dog_stage_value that have the value "None" to nulls using replace and np.nan
2. I changed the data type in dog_stage_value from object to category using astype.
3. I removed from the dataframe the values that are retweets using masking and .isnull on the column retweeted_status_id'
4. I removed values in the column name that are wrong like 'a', 'the' and 'an' using replace and np.nan
5. I changed the datatype of the timestamp column to date using to_datetime from pandas.

6. For the Albanian 3 1/2 legged Episcopalian, I changed the numerator and denominator values to the correct values using str.contains, masking and loc.
7. I removed tweets that don't have images using masking and str.contains("http")
8. I removed duplicated timestamp values, dropping the duplicated column of timestamp after joining tables.¶