WRANGLE ACT PROJECT

My wrangling process included the following steps:

Data Gathering

The process of gathering data from 3 different sources:

1. A CSV file provided by udacity for the twitter archive
2. The Images data table gathered programmatically by sending a request to the file host
3. The twitter feed gathered through the tweepy twitter API

Data Wrangling

The second part was to inspect that data visually and programmatically using PANDAS which revealed several data issues including:

ISSUE	SOLUTION
QUALITY	
Retweets detected	Removed programmatically from the table
Wrong dog names	Replaced by None
Tweet_Id should be converted to string	Converted through pandas .astype()
timestamp should be converted to a date time object	Converted through pandas .astype() and correct Regex
Incorrect rating numerators and denumerators	Correct values are extracted from tweet text
Dog types need to have the same format (Capitalized or smalled) and underscores removed	Corrected
Images that are idientifed as not dogs should be removed	Rows removed through conditional looping
ID COLUMN NAME SHOULD CHANGE TO TWEET_ID	Changed in twitter feed table
TIDINESS	
STAGE COLUMN SHOULD COMBINE THE 4 EXISTING DOG STAGES INTO ONE COLUMN	The 4 columns were combined in a single column indicating the dog stage
MERGING THE DATASETS SO THAT IT IS EASIER TO ACCESS THE DATA FOR LATER ANALYSIS	The data sets were merged based on the tweet_id