Data Wrangling Report

Project Overview
Exploring twitter archive of twitter user @dog_rates also known as WeRateDogs. WeRateDogs is a twitter account known for rating peoples dogs and tweeting humorous comments about them.

Project Objective
Wrangle and Store WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Step 1: Gather
The tweet archive was provided by Udacity and downloaded manually. Additional data queried from twitters API using tweet ids present in the archive dataset, was downloaded manually and read line by line into a pandas dataframe. An image prediction file from a neural network that classifies dog breeds contains a table of the top three image predictions of dogs and was downloaded programmatically from Udacity's servers with the Requests library.

Step 2 and 3: Assess and Clean
To assess each dataset for errors, visual and programmatic assessment was carried out. The table below reports the number of quality and tidiness issues observed, and how it is resolved.

Quality

Dataset	Observation	Solution
twitter_archive	Missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp variables.	Since both issues can be solved simultaneously, columns were dropped.
	Irrelevant columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)	
	Incorrect data in name column ('None', 'a')	Converted 'None', and 'a' to null
	Url string format in source	Extracted accurate source from url
	Null in stage represented as 'None'	Converted 'None' to null
	Erroneous data type in timestamp	Changed data type to datetime
	Some tweets do not have images	Deleted rows with null value in img_num column

	Some tweets are retweets	Deleted rows in text begining with “RT”
	Some tweets have multiple dog stage, with some describing two dogs	Changed to the correct stage after referring to the corresponding data in text and jpg_url columns, and added a separator for stages describing two dogs
image_predictions	Not all values in p1, p2, and p3 are dog breeds	Deleted rows that indicated false in prediction
	Lower case and upper case sometimes in p1, p2, amd p3	Changed values to lower case
tweet_detail	Missing records (2354 non-null instead of 2356)	After merging with twitter archive, null values were found in archive as retweets, and deleted.

Tidiness

Dataset	Observation	Solution
twitter_archive	doggo, floofer, pupper, puppo should be observations in one column dog_stage	Concatenated columns into one column, stage
image_predictions	Three columns represented as p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog, p3_dog	Reshaped to three columns as prediction, confidence_lvl, and breed.
	tweet_id should be present in twitter_archive table	Merged three tables to create master dataframe and saved as twitter_archive_master
tweet_detail	favorites, and retweets should be added to the twitter_archive table	
	tweet_id should be present in twitter_archive table	