Data Wrangling Report

Project Objectives

The project main objectives were:
• Perform data wrangling (gathering, assessing and cleaning) on the provided sources of data.
• Store, analyze, and visualize the wrangled data.
• Reporting on

1. data wrangling efforts.
2. data analyses and visualizations.

Gathering

In this phase three data sources are loaded in different ways then loaded into pandas DataFrame:

- The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archive-enhanced.csv')

• The tweet image predictions ('image_predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.

• Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Assessing and Cleaning

While Assessing data, a number of Issues were observed. In the table below representing the issues along with actions taken in the cleaning Step.

Quality

DataFrame	Issue	Solution
archive_df	Columns (doggo, floofer, pupper, puppo) has None for missing values.	Replace None values with np.nan
	expanded_urls has NaN values.	Remove NaN entries
	rating_numerator column has incorrect values	Convert it to float and extract the value correctly from the text using RegEx
	rating_denominator column has values less than 10 and values more than 10 for ratings more than one dog.	Investigate the values that can be fixed and remove the others.
	text column has the link for the tweets and ratings at the end we can remove it.	Remove ratings and links using regex`'(.+(?=\s\d+/\d+\s))'`.
	timestamp is a string instead of datetime.	Convert dtype to datetime.
	We are interested in the tweet only not the retweet or reply.	Remove retweets and replies.
	Has non-dog tweets.	Remove any non-dog related tweets.
	name has invalid values.	Replacing the invalid values with np.nan.
api_df	id column needs to be renamed as the other 2 datasets.	Rename id column into tweet_id.
	Has unnecessary columns.	Remove the unnecessary columns.
img_df	img_num column is useless.	Remove the column.

Tidiness

DataFrame	Issue	Solution
archive_df	Columns (doggo, floofer, pupper, puppo) are all about the same data. (dog_stage)	Combine them into one column.
img_df	Columns (p1, p2, p3), (p1_conf, p2_conf, p3_conf), (p1_dog, p2_dog, p3_dog) has non-descriptive names and each line is about the same data.	taking the highest confident prediction as a dog tweet otherwise np.nan.
All Datasets	All data is related but separated into 3 datasets.	Combine the 3 datasets into only one.

Output

A combined dataset with all information stored in sqlite database (twitter_archive_master).