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.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:
• The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.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.

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Dataset	**Observation**	**Solution**
Twitter Archive	**Quality** ● some of the column names are not meaningful (like: retweet, reply to the original tweet, ..) ● timestamp column is str instead of datetime **Tidyness (structure)** ● more than one stage is filled for a particular dog ● "source" and "expanded_urls" have several informations inside them. ● doggo, floofer, pupper and puppo refer to the same dog stage. ● rating_numerators should be of type float and	- Replaced Non values with np.nan. - Removed the rating score and tweet link from the tweets text column using RegEx and pandas exctract method. - Converted timestamp to datedime data type using pandas to_datetime function. - Removed retweets rows from data. - Removed replys rows from data. - Extracted the rating score correctly and converted it to float. - Removed any rows with denominator more than 10.

	are not always correctly extracted.	- Removed rows with missing expanded urls as they are not valid data. - Replaced None and unvalid names with np.nan.
Image Predictions	**Quality¶** ● there are 2075 tweet id, and the archive dataset has a total of 2356 ids which means 281 IDs are missing. ● columns name are not the best thing. ● p1, p2, and p3 contain underscores instead of spaces in the labels. ● img_num is not needed.	- Removed other columns - Renamed it to match the other 2 datasets
JSON File	**Quality** ● the original twitter_arch has 2356 tweet_id and JSON file 2354 (number of missing IDs = 2356-2354 = 2)	

All	All datasets should be combined into 1 dataset only	Combined all the 3 datasets into one pndas df

Result

A combined data set with all needed information was stored in CSV file called "twitter_archive_master_new.csv"