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.

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.

Quality

Dataset	Observation	Solution
df_arch	Columns (`doggo`, `floofer`, `pupper`, `puppo`) has `None` for missing values.	Replaced Non values with np.nan
	`text` column has the link for the tweets and ratings at the end we can remove it.	Removed the rating score and tweet link from the tweets text column using RegEx and pandas exctract method.

	`timestamp` is `str` instead of `datetime`	Converted timestamp to datedime data type using pandas to_datetime function.
	We are interested in the tweet ONLY not the retweet there for we should remove those from the table.	Removed retweets rows from data.
	We are interested in the tweet ONLY not the reply to the original tweet there for we should remove those from the table.	Removed replys rows from data.
	The `rating_numerator` column should of type `float` and also it should be correctly extracted.	Extracted the rating score correctly and converted it to float
	`rating_denominator` has values less than 10 and values more than 10 for ratings more than one dog.	Removed any rows with denominator more than 10
	`expanded_urls` has NaN values	Removed rows with missing expanded urls as they are not valid data
	`name` column have None instead of NaN and too many unvalid values.	Replaced None and unvalid names with np.nan
df_api	`id` column name different than the other 2 data sets.	Renamed it to match the other 2 datasets

Tidiness

Dataset	Observation	Solution
df_arch	`doggo`, `floofer`, `pupper`, `puppo` columns are all about the same things, a kind of dog personality.	Created one colum dog_stage and removed the 4 columns
df_pred	`img_num` useless.	Removed it
	Just 3 columns needed `id`, `retweet_count`, `favorite_count`	Removed other columns
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 a sqlite data base.