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 'twitterarchiveenhanced.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
twitter_archive	Columns (doggo, floofer, pupper, puppo) has None for missing values.	Replaced Non values with np.nan.
twitter_archive	timestamp is str instead of datetime	Converted timestamp to datetime data type using pandas to_datetime function.
twitter_archive	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.
twitter_archive	We are interested in the tweet ONLY not the retweet there for we should remove those from the table	Removed retweets rows from data.

twitter_archive	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.
twitter_archive	tweet_id is integer instead of string	Converted tweet_id to str data type using pandas astype function.
twitter_archive	Some dog names are invalid('a','an','None')	Replace them with np.nan Then remove them.
Twitter_API	id column name different than the other 2 data sets	Renamed it to match the other 2 datasets.

Tidiness

Dataset	Observation	Solution
twitter_archive	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
image_prediction	Missing photos for some IDs	Removed it
	Just 3 columns needed id, retweet_count, favorite_count	Removed other columns
	Some P names start with uppercase and other with lowercase	Replace P name that has lowercase with uppercase
all	All datasets should be combined into 1 dataset only	Combined all the 3 datasets into one pandas df

Result

A combined data set with all needed information was stored `twitter_archive_master.csv`