Data Wrangling Report

Project objectives

The project's main objectives were:

- Perform data wrangling (gathering, assessing and cleaning) on provided data sources.
- Store, analyze, and visualize the wrangled data.
- Reporting number (1) data wrangling efforts and (2) data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas' data frames:

- Manually downloaded the WeRateDogs Twitter archive ('twitter-archive-enhanced.csv').
- The tweet image predictions dataset ('image-predictions.tsv') was downloaded programmatically using the Requests library from a provided URL.
- Read the resulting data from twitter_api.py (tweet_json.txt file) line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favorite count."

Steps 2 and 3: Assessing and Cleaning Data

several observations were made while working with data. The table below shows the data wrangling process in detail.

Quality

Dataset	Observation	Solution
image_predictions	Integer instead of string datatype in "tweet_id" column.	This was converted to string data type.
	Underscores instead of space in "p1", "p2" and "p3" columns value.	The underscores were removed and replaced with a white space.
	The first character in every word in p1, p2, p3 columns are in lower case.	All values were edited to start with a capital letter.
	There are 2075 entries in "image_predictions" while 2356 in "twitter_archive", resulting to 281 missing IDs.	This was not changed as we will need all data with image in our data analysis and visualization.

twitter_archive	Integer instead of string datatype in “tweet_id” column.	This was converted to string data type.
	The "timestamp" column is not a datetime datatype.	This was converted to datetime data type.
	The dog names are not standardized.	Faulty dog names were replaced with None.
	Retweets are present in the dataset.	Columns that are retweets were dropped.
	Most of the text in the "text" column have links in the end.	Links were removed from text.
tweet_json	There are 2354 entries in "tweet_json" while 2356 in "twitter_archive", resulting to 2 missing IDs.	This issue was solved after merging all datasets to one.

Tidiness

Dataset	Observation	Solution
twitter_archive	"source" and "expanded_urls" column have several information inside them.	These columns were dropped since it does not much impact in the dataset.
	Columns "doggo", "floofer", "pupper", and "puppo" refers to the same unit measurement.	These columns were merged in to one column called “dog_stage”.
	Some dogs are classified in more than a stage.	No changes were made to this because a large proportion of the dog are not classified at all. Thus, wound not affect our dataset.