Wrangle Report

This report intends to outline the project's progress throughout the project wrangle phase. The information is generated from a Twitter account that rates various breeds of dogs. A scale from 1 to 10 is used to rate the submissions.

Three steps have to be finished in order for the full analysis of this data to be completed.

1. Gathering Data

- First dataset was “twitter-arhive-enhanced.csv” which I loaded from the project provided material.
- The second dataset was the “image prediction” which I downloaded programmatically using the *requestsAPI* the and loaded it programmatically.
- The final dataset was scraped from twitter using the *tweepyAPI* which requires from You to make a Twitter developer account to access the data, and you must be given permission to use it; otherwise, you won’t be able to use it. The file containing the scraped data is named as “tweet-json.txt”.

2. Assessing Data

After gathering all the required datasets for the projects, the following steps were taken precisely to asses the quality of the three datasets

- Examined the datasets for tidiness or quality problems.
- .info() method was used to obtain a summary of the dataset.
- duplicate rows and null values were checked.

- A specific column of interest, such as rating numerator and rating denominator, was examined for value counts.

3. Cleaning Data

Following a careful examination of the data during the assessment step, the following points were cleaned:

- Only kept the original ratings (no retweets) for tweets with images.
- Eliminated rating denominators that weren't equal to 10.
- discarded a few retweet-related columns that were unnecessary.
- extracted the text from the source column and removed the HTML tag "a"
- Name column naming is uneven and was cleaned to be more realistic.
- Fixed the Incorrect object datatype for the timestamp column
- Replace "None" in columns (doggo, floofer, pupper, and puppo) with np.nan
- combine the columns for doggo, floofer, pupper, and puppo into one.
- Make a master dataframe out of all the datasets.