Wrangle Report

This report aims to show the step taken in the wrangling phase of the WeRateDog project. The data are generated from a twitter account(WeRateDogs) that gives rating to different types of Dogs. The rating are given on a scale of 1-10 and often feature comedic comments an jokes.

In the full analysis of this data, three stages had to be completed to come to completion, namely;

- ✓ Gathering Data
- ✓ Assessing Data
- ✓ Cleaning Data

Gathering Data

Three dataset are used in the project, the first of which is downloaded directly from the course material provided **"twitter-arhive-enhanced.csv"**. The second dataset was downloaded programmatically with the request library, the dataset is the **image prediction**. The third dataset is scrapped from twitter via twitter api using the tweepy library, to access the data you need a twitter developer account and must be granted access to use it else the file containing an already scrapped data is provided name **"tweet-json.txt"**

Assessing Data

The three dataset after completely gathered, the following were steps were taken as shown below:

- ✓ Checked the head, tail and random sampling of the dataset to spot quality or tidiness issues.
- ✓ Used the .info() method to get an overview of the dataset
- ✓ Checked for duplicated rows as well as null values
- ✓ Checked for value counts of specific column of interest.

Cleaning Data

After a thorough look through the data in the assessment stage the following observations were cleaned:

- ✓ Removed the html tag <a> and extracted only the text from the source column
- ✓ Changed the timestamp datatype to datetime from an object datatype
- ✓ Filtered out rating_denominator that were not equal to 10
- ✓ Replace "None" in columns (doggo, floofer, pupper, and puppo) with np.nan
- ✓ Remove link for the tweets and ratings at the end of the text column
- ✓ Invalid values that starts with lower case and None' are replaced np.nan in the name field
- ✓ expanded_urls columns has NaN values
- ✓ Created a new column named "dog_stage" and grouped the growth stages
- ✓ Merge all dataset into a master dataframe
- ✓ Changed the data_type of the tweet_id column to string from integer