WRANGLING REPORT

The word wrangle means to round up, herd and take charge of livestock. In this report, I will discuss the wrangling activities I carried out for the WeRateDogs twitter archive.

The wrangling process consists of:

- Gathering: - Gathering the data from various sources.
- Assessing: - checking the data for any issues with its quality or structure.
- Cleaning: - cleaning the data by removing the items causing the data to be messy and dirty.

I will be using the pandas package in python to perform my data wrangling and it will be done using Jupyter Notebook.

Gathering

- The WeRateDogs twitter archive: - I downloaded this dataset from the Udacity server as a **csv** file which I then uploaded into python using the pd.read_csv function.
- The image prediction file: - This file was downloaded programmatically using the request library in python from a URL and stored in a folder as a **tsv** file and then uploaded into python using the pd.read_csv function .
- The twitter API dataset: - The tweepy library was used to query the twitter API and the entire json data for each tweet was downloaded.

Assessing

I assessed the dataset both visually and programmatically and noted a few quality and tidiness issues which were detailed in my work. A few of the Quality issues I observed were:

- The tweet_id column should have the same name in all Dataframes.
- The are redundant retweet rows.
- The are redundant rows in the in_reply_to_status_id column.
- There are some columns in the t_archive dataset that are not relevant to the analysis.
- Some of the values in the rating_numerator column were not extracted properly.
- The timestamp column in the t_archive dataset has the wrong datatype.The tweet_id column in the t_archive dataset also has the wrong datatype.
- The dogs without names in the t_archive dataset being labeled as 'a' or 'an' or 'the'instead of None.
- Some values in the rating_denominator column in the t_archive dataset do not equal 10.

As for tidiness issues observed:

- I observed that the four dog stages (doggo,floofer,pupper,puppo) in t_archive dataset were spread across 4 columns when they should all be under one column.
- The retweet_count,the favorite_count and jpg_url columns are not in the t_archive dataset.

Cleaning

In the cleaning section, I cleaned the issues I had raised during the assessment phase. I began by creating a copy of my datasets. A few of the cleaning that was done are as follows:

- Rename the id column in the twt_js dataset to tweet_id.
- Find the index for the retweet rows and then drop them.
- Find the index for the redundant rows in the in_reply_to_status_id column and then drop them.
- Drop the 'source' column and the 'expanded_urls' column.
- Changing the datatype of the Timestamp column to datetime datatype.
- Changing the datatype of the tweet_id column to str datatype.
- Replacing ‘a’, ‘an’ and ‘the’ with ‘None’.
- Dropping the rows where rating_denominator does not equal 10.
- Putting all four dog stages under one column named ‘Dog_stages’.
- Merging the t_archive, img_prd and the twt_js datasets using the shared column of ‘tweet_id’.

The cleaned data was stored as “**twitter_archive_master.csv**” file. The image prediction and the twitter API datasets were not saved because they were not cleaned.