@WeRateDogs Data Wrangling Report

Project Objectives

- Gathering, Assessing and Cleaning the data.
- Analysis and visualization of the data.
- Giving detailed reports of the steps taken to achieve the above mentioned

Step 1: Data Gathering

I successfully gathered three datasets;

The Archived tweets file from @WeRateDogs('twitter-archive-enhanced.csv'): this was supplied on-hand by Udacity and I only needed to download manually. Then I read it into a pandas dataframe.

The images prediction data('images-prediction.tsv'): the URL to this file was given. I had to download it programmatically using the "Requests" library from the given URL. The data was then read into pandas dataframe.

The API gathered data('tweet_json.txt'): I had to query twitter's API to get the data needed. Data collected include; tweet_id, retweet_count, favorite_count and the followers_count. Then the data was stored and read into pandas dataframe.

Step 2: Assessing the Data

A thorough approach was taken in assessing the data. It was done in two phases

Visual Assessment using Microsoft Excel and

Programmatic Assessment using Pandas methods

From the assessment, the following observations and actions taken are recorded:

Quality Issues

Dataset	Column	Observation	Solution
twt_df	**tweet_id**	Datatype is string, should be float	Changed datatype to float
	in_reply_to_status_id	datatype should be string not float	changed datatype to string
	in_reply_to_user_id	datatype should be string not float	changed datatype to string
	timestamp	Datatype should be datetime not string	changed datatype to datetime format

	rating_denominator	Twenty three records with a denominator not equal to 10	No action taken. The denominators need not be equal to ten.
	Rating_numerator	Wrong ratings of decimal values	Extracted correct ratings from the text column. Created New column taking the ratio of numerator to denominator.
	name	Invalid names recorded	Changed all invalid names to null
	name	Missing names represented ny 'None"	converted all 'None' to 'NaN' using np.NaN
	name	Some names starting with small letters	Capitalized all names
	name	"O" instead of "O'Malley"	Changed name to "O'Malley".
	Doggo, Floofer, Pupper, Puppo	Missing values not represented as NaN	Used the np.NaN to convert all missing values.
	retweeted_status_id, retweeted_status_user_id, retweeted_status_time_stamp	Irrelevant columns	Dropped the columns
		Records of retweets and replies	Records dropped from the dataset

Tidiness Issues

Dataset	Column	Observation	Solution
twt_df	**Doggo, Floofer, Pupper, Puppo**	Some records have more than one dog stage	No action taken. These records have more than one dog in picture as discovered through the text column. Concatenated multiple dog stages into one column.
	Doggo, Floofer, Pupper, Puppo	Should be under one column called dog stage	Created a new column called dog_stage and filled appropriately. In the case of more than one dog stage, it fills the stages and joins them e.g 'doggo, floofer'.
The three Datasets		Should be one dataset, since the data in each all form part of a observational data(tweets)	Merged the three datasets

And finally, we have the datasets all tidy and cleaned up.