Data Wrangling Report

This report summarizes the data assessment and cleaning efforts in order to achieve "tidy" data for the Twitter WeRateDogs data which further facilitates analyzing and gaining insights from the data. First, the data was gathered from three different sources with three different formats. Data was queried, downloaded, and loaded into the workspace using the tweepy and pandas libraries.

Each of the gathered datasets were assessed separately. The table below summarizes the identified issues and its proposed fix. More details about the define, code, test steps can be found in the wrangle_act.ipynb file.

Finally, the cleaned data has a total of 1873 observations. Some missing values are present in the data to avoid losing valuable information from the large number of observations of certain variables if all the missing values are removed. Therefore, the data is queried as per the variables required for answering each question.

Type	Data	#	Issue	Fix
Quality	WeRateDogs Twitter Archive	1	Dog names have missing "None" values rather than NaN.	Replace None with npnan in columns ('name', 'doggo', 'floofer', 'pupper', 'puppo')
		2	Some expanded_urls are repeated multiple times (split with a comma).	Split expanded_urls strings by comma and replace with only the string before the first comma
		3	Data types of id columns (tweet_id, in_reply_to_status_id, in_reply_to_user_id,	Change the datatypes of id columns (tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id) to string
		4	Data types of the timestamp columns (timestamp, retweeted_status_timestamp) are strings rather than datetime.	Change the datatypes of the timestamp columns (timestamp, retweeted_status_timestamp) from strings to datetime
		5	The data has 181 retweets while we are interested in original ratings only	Deleting retweets and in_reply_to tweets
		6	The data has 78 in reply to tweets while we are interested in original ratings only.	
		10	Rating_numerator and rating_denominators should be floats rather than integers	Change the datatype of the rating_numerator and rating_denominator columns from integers to floats

		13	After removing "in_reply_to" and "retweets" ratings, the columns related to them (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) will become redundant	Drop in_reply_to and retweeted columns ['in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp']
		16	several values that are not dog names, like 'a', 'the', 'such', etc.	Use lowercase characters to clean up dog name column.
		18	rating_numerator, and doneminator are not extracted correctly	pick the decimals in the ratings using a regex
	Tweet image prediction	7	Data type of the tweet_id column is integer rather than a string.	Change the datatype of the "tweet_id"column from integer to a string
		11, 12	tweets without a jpg URL or a rating	Drop tweets that do not have images or ratings
		14	img_num column and the other columns related to the second and third prediction are redundant.	Drop img_num column and other columns related to the second and third prediction ['img_num','p2','p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog']
	Twitter API	8	the id column name does not match the tweet_id column name in the other datasets	Change the "id" column name to "tweet_id" to match the other datasets
		9	the datatype of "id" column is integer rather than string	Change the datatype of the "tweet_id"column from integer to a string
		15	the data has 31 columns but only three columns are needed	Remove all columns except ["tweet_id", "retweet_count", "favorite_count"]
		17	retweet_count and favorite_count should be integers, not floats.	Change the datatype of the retweet_count and favorite_count columns from floats to integers
Tidiness	WeRateDogs Twitter	1	Columns doggo, floofer, pupper, and puppo represent the same variable type which is dog breed and should be gathered into a single column	Create dog_stage column that gathers doggo, floofer, pupper, and puppo columns into a single column.
	General	2	merging individual datasets of data into a single dataframe	Merging the cleaned datasets into a single dataframe