WeRateDogs®

twitter.com/dog_rates

DATA WRANGLING REPORT
February, 2021

PROJECT OBJECTIVES:

WeRateDogs (under the Twitter handle @dog_rates) is an account focused on rating dogs often with photos, commentary and a rating, done in a light, entertaining spirit.

Project focuses on:

- **data wrangling** focusing on tidiness issues and data quality from raw data files
- storing, **analyzing and visualizing the data**
- producing an **internal report** on data wrangling efforts and a **public facing report** on data analyses and visualizations.

DATA SOURCES:

1. **`twitter_archive`** : WeRateDog Twitter archive data provided file to be imported via pandas. This contains basic tweet data for all tweets as of August 1, 2007.
2. **`image_predictions`** : Top 3 image predictions of each tweet (goal of predicting the breed of dog shown) according to neural network. The file ('image_predictions.tsv') is hosted on Udacity's server and downloaded using the request library via provided url.
3. **`tweet_data`** : Using tweet IDs in 'twitter_archive', query Twitter API for each tweet's JSON data using tweepy library and store each tweet's entire set of JSON data in a file 'tweet_json.txt'.

ASSESSMENT OF DATA/ CLEANING PLAN:

QUALITY ASSESSMENT

Dataset	Observation	Solution/Cleaning
`twitter_archive`	Missing data found in the following columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'.	Since the columns mention contain at most data for 7% of the dataset - these columns will be dropped since not enough data is available or can be available to use.
	The datatype of 'timestamp' is not formatted as datetime data.	The datatype will be changed accordingly.
	Column 'expanded_urls' have missing values indicating no images.	There are 59 tweets that do not contain a picture indicated by no value in the expanded_url column. Since these tweets contain information outside the scope of the analysis these lines will be removed from the dataset.
	First letter of dog names are inconsistently capitalized.	The first letter of all dog names will be capitalized for consistency.
	'source' columns contains heavy residual text from html code.	The content will be consolidated to a more concise description based on current values.
	There are a few ratings with a denominator other than 10.	Based on value_counts for the denominator and given information about the Twitter account it appears that there are some lines with a value besides 10. This includes 18 rows. The text will be examined of these rows closer to determine what approach should be done.

QUALITY ASSESSMENT *(continued)*

image_predictions	The prediction of dog breeds (or object) have inconsistent capitalization.	For consistency, the values in the prediction columns will all be lower case. Also the underscore character ('_') currently used to separate more than one word will be replaced with a space.
	Column headers "p_1" etc is a little ambiguous to its contents.	"p_1" (and similar) will be renamed "prediction_1" which will help clarify the nature of the contents of the columns.

TIDINESS ASSESSMENT

Dataset	Observation	Solution/Cleaning
twitter_archive	The columns in `twitter_archive` should be reported as a value under one column: 'doggo', 'floofer', 'pupper', and 'puppo'	Merge columns under one heading “stage” since only one column among the 4 is used
tweet_data & twitter_archive	Tweet_data contains additional details of each tweet	tweet_data and twitter_archive will be merged into one table