DATA WRANGLING REPORT – Jun 13, 2022.

We Rate Dogs Analysis

Summary

This report outlines the data wrangling process as it relates to gathering, assessing, and cleaning data from relevant sources to arrive at meaningful insights from the activity of the WeRateDogs twitter account. After wrangling, the relevant datasets were combined into a single dataframe, then stored locally in preparation for further analysis.

Gathering Data

WeRateDogs downloaded their Twitter archive and shared it with Udacity for use in this project. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The archive data, twitter_archive_enhanced.csv, was downloaded directly from the Udacity classroom. Additional data was gathered from two other sources:

- An image prediction file, image_predictions.tsv, containing information about the breed of dogs present in the archive. This file was programmatically downloaded with the requests library.
- Each tweet json data was queried from the Twitter API using the Tweepy library, then locally stored to tweet_info.json. Useful information such as retweets, likes and hashtags were programmatically pulled from this file.

Gathered data was read into three dataframes: *wrd_archive, img_predictions*, and *tweet_json_info*.

Assessing Data

The assessment process involved examining the data visually and programmatically to identify possible issues with data quality and tidiness. Notable observations included quality issues like:

- Unwanted records that were not original posts, like retweets and replies.
- Erroneous data types for the *tweet_id* and *timestamp* columns.
- Unexpectedly high ratings and in some cases, wrong ratings extracted from tweet text.
- Inconsistent ratings needing to be standardized.
- Occasional classification of dogs into the wrong stages.
- Invalid dog names: these entries were typically formatted in lowercase.
- Non-dog related tweets found in the archive.
- Numeric information like retweets and likes count were stored as string or object types.
- Unequal number of records between the three dataframes.

Notable tidiness issues were also observed:

- Variables such as predictions, confidence and accuracy were scattered across several columns.
- Dog stage information stored across four columns rather than in a single variable.
- The *Source* column contained long and unnecessary information (HTML tags and source).
- The *text* column contained two variables: tweet url and tweet text.
- Where present, multiple hashtags were stored in python lists.

All quality and tidiness issues were clearly documented to aid proper cleaning.

Cleaning Data

The cleaning process was conducted on copies of the three dataframes with the define-code-test framework utilized throughout. Notable cleaning actions involved:

- Fixing erroneous datatypes with the *.astype()* and *pd.to_datetime()* methods.
- Removing unwanted records for retweeted posts and replies.
- Using regular expressions to replace unusual dog names with *None*.
- Extracting the actual tweet source from the *source* column.
- Eliminating all unwanted columns.
- Standardizing all ratings to a scale of 10.
- Separating the tweet url from the tweet text.
- Splitting multiple hashtags into individual records using *df.explode()* method.
- Placing all the dog stages into a single column.
- Selecting the most confident prediction for each breed, storing the prediction value and confidence in respective columns, then dropping redundant columns.

Cleaned copies of all dataframes were merged into a master dataframe, *twitter_archive_master.csv*, to ensure that only common records are retained; setting the stage for further analysis and visualizations.