DATA WRANGLING REPORT – June 15, 2022.

We Rate Dogs Analysis

Summary

This report outlines the data wrangling process as it relates to gathering, assessing, and cleaning WeRateDogs twitter data, collected in fragments from different sources. 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. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. This 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 on the breed of dogs present in the archive: The file was programmatically downloaded with the **requests** library.
- Twitter API: Each tweet json data was queried using the **tweepy** library, then stored locally to tweet_json.txt. Information on retweets, likes and hashtags were 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.
- Missing entries in the **expanded URL** column.
- Unequal number of records across the three dataframes.

Notable tidiness issues were also observed:

- Variables such as **predictions**, **confidence** and **accuracy** were scattered across several columns.
- Dog stage information was 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, **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 **df.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, then stored into **twitter_archive_master.csv**. The master dataframe comprises **1961 rows** and **12 columns.**

Conclusion and Recommendation

Data wrangling is an iterative process, and one might find more issues with the datasets than I have originally identified. At the moment, this cleaning has sufficiently set the stage for further analysis and visualizations.