WeRateDogs – Data Wrangling Report

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. It has over 4 million followers and has received international media coverage. In this project, data about WeRateDogs Tweets are collected from a number of sources, assessed, and then cleaned for final analysis. This report describes how the data were collected and the measures taken to clean the data.

Data Gathering

The raw data were gathered from three different sources:

1. Twitter archive data – manually downloaded from the link provided in "Project Details" and then uploaded to the project workspace. Then the data was loaded into a DataFrame.
2. Image predictions data – programmatically downloaded from a link provided in "Project Details." The data were saved onto the project workspace and then loaded into a DataFrame.
3. Additional data through Twitter AP – consumer key and access token were obtained through Twitter developer account. A list of Tweet IDs from the Twitter archive file was used to programmatically retrieve extra data in JSON format from Twitter API. Specific data ("favorite_count" and "retweet_count") were extracted and saved as a file.

Data Assessment and Cleaning

Twitter Archive File

- No duplicated rows were found.
- As stated in Project Motivation, only original ratings are considered for final analysis. Therefore, rows that are retweets or replies were dropped from the dataset. That is, only the rows with "Null" cell in reply- and retweeted-related columns were kept.
- Reply- and retweeted-related columns were also removed as they are no longer relevant.
- Entries in "timestamp" column were converted to datetime format.
- Six Tweets were found with wrong denominator due to extraction error. The texts of these Tweets were read and the ratings changed manually.
- The Tweets with 10 as rating denominator but with deviant values (higher than 14 in this case) as rating numerator were dropped from the dataset.
- Three Tweets are without "expanded_urls" and can't be found in image predictions file. Only the Tweets with URL (not null in "expanded_urls" column) were kept.
- Deviant values in the "name" column were replaced with "None."

- Four columns of dog stage were melted into a “dog_stage” column. Value “none” was assigned to the Tweets without dog stage. If a Tweet has two stage, the one with lower count was kept.

Extra Data through Twitter API File

- The DataFrame with extra data (“retweet_count” and “favorite_count”) through Twitter API was merged with Twitter archive DataFrame.
- Tweets without “retweet_count” and “favorite_count” in the compiled DataFrame were removed.

Image Predictions File

- Tweets with a dog breed in all three predictions were selected, and the prediction with highest confidence was to “breed” and “confidence” columns.
- “breed” and “confidence” columns were to Twitter archive DataFrame.

Final Cleanup

- Irrelevant columns in the Twitter archive DataFrame were dropped.