WeRateDogs: Data Wrangling Process

Introduction

This report details the data wrangling process for three data sets, the tweet_archive, extra tweet data, and the image_predictions which were gathered from various sources. The tweet_archive was provided as a CSV file that contains basic information about tweets, including the text, timestamp, URL, and some extra information extracted from the tweet text by the author such as ratings, dog name, and dog stage. The image_predictions however was a TSV file hosted on a server to be downloaded programmatically, it contains three image predictions and confidence levels of dog breeds in the tweets. The third was the favorite count and retweet count of tweets present in the tweet archive, this was to be gathered by querying the Twitter API for each tweet's JSON data.

Data Gathering

The tweet archive was read as a local file ('twitter-archive-enhanced') into a Pandas data frame. The image prediction TSV file was requested from the Udacity server and loaded into a Pandas data frame as well. The extra tweet data was gathered using python's **Tweepy** library to query each tweet's JSON data and store each tweet's entire set of JSON data in a file called 'tweet_json.txt' line by line, this file is then read line by line into a Pandas data frame storing only the tweet id, favorite count and retweet count.

Assessing Data

After gathering the data, an initial assessment was conducted visually and programmatically to identify any quality or tidiness issues that needed to be addressed. The following issues were identified:

Quality issues - tweet_archive table

- Tweets include replies and retweets
- Nulls represented as 'None' in dog name and dog stages columns
- The name column contains odd characters ('a')
- Missing values (name, expanded url, and dog stage)
- Rating numerator of 0 and only the digit after the decimals were recorded for floats (75, 27, 26)
- Inconsistent rating denominator
- Source column values are wrapped in HTML syntax
- Erroneous datatype (timestamp)
- Not all tweets have image predictions

Quality issues - image_predictions table

- Inconsistent case use of dog predictions (upper and lower case)
- False predictions (most confident/all prediction has a false breed of dog)

Tidiness issues

- One variable (dog stage) is split into four columns (doggo, floofer, pupper, puppo)
- tweet_extra should be part of tweet_archive
- Most confident predictions not highlighted in the image_predictions table

- Redundant columns in tweet_archive table (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Redundant columns in image_predictions table (rating_denominator)
- Add the most confident prediction column from the image_prediictions to the tweet_archive

Cleaning Data

To address the issues identified during the assessment phase, the following cleaning steps were taken:

- Replies and retweets were removed from the tweet_archive table.
- Null values represented as "None" in the dog name and dog stages columns were replaced with NaN.
- Some of the names with odd characters ‘a’ were extracted from the text and others were replaced with ‘unknown’
- Missing values in the name expanded url, and dog stage columns were replaced with ‘Unknown’.
- Errors in the rating_numerator and rating_denominator columns were corrected and standardizing the ratings to a denominator of 10
- HTML syntax was removed from the source column.
- The erroneous timestamp datatype was corrected.
- False predictions were removed and an Iterated assessment was done after excluding images with false predictions
- Inconsistent case use of dog predictions was standardized
- One variable (dog stage) split into four columns was addressed and the redundant columns were dropped
- tweet_extra was merged with the tweet_archive table
- The most confident prediction was highlighted in the image_prediction table and added to the tweet_archive table

Conclusion

The data wrangling process for the tweet_archive and image_predictions table was a multi-step process that required careful attention to detail. Several quality and tidiness issues were identified and addressed through a combination of methods.