Udacity Data Wrangle Project – We Rate Dogs – Wrangle Report

In this report, I outline the effort made to gather, assess, and clean the data required for analysis of the WeRateDog twitter archive.

Data Gathering

I gathered data from three different sources, stored in separated files:

1. WeRateDogs twitter enhanced archive downloaded manually from udacity servers
2. Image predictions file downloaded programmatically from udacity servers
3. Tweet JSON file downloaded manually from udacity servers

Data Assessment

I began the assessment by viewing information on the archive data frame first then image predictions, and JSON data frame identifying some quality and tidiness issues like:

Quality

- Tweets with expanded_urls indicating 59 tweets with missing data.
- There are 181 retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp).
- There are 78 reply tweets (in_reply_to_status_id, in_reply_to_user_id).
- The name column has 109 invalid name values. Invalid words are all lowercase; all valid names start with an uppercase letter.
- rating_denominator has some inconsistent values it shows as high as 170 respectively. Ignoring replies and retweets, there are 17 tweets with rating_denominator not equal to 10.
- There are 28 tweets with rating_numerator >= 15. The max value is 1776, When we only look at tweets with a rating_denominator of 10, there are 12 tweets with rating_numerator >= 15.

- The source column can be simplified by extracting string from <a> tag.
- There are 2075 image predictions, 281 less than the number of tweets in the archive
- Erroneous datatypes (timestamp, source, dog_stage, tweet_id)

Tidiness

- 1. There are 4 columns for dog stages (doggo, floofer, pupper, puppo), which doesn't conform to the rules of "tidy data".
- 2. We are only interested in "original tweets", no "retweets"; this data is stored in the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- 3. Reply tweets are also not "original tweets", this data is stored in the columns in_reply_to_status_id, in_reply_to_user_id.
- 4. rating_denominators is no longer needed as values are the same (10).
- 5. The dog breed prediction with the highest confidence level can be combined with the archive table as the twitter table contains information that is all about the dog in the tweet.
- 6. The json_data table should be combined with the archive table.

Data Cleaning

I began the cleaning by creating a copy of archive and image predictions data frames then I handled every single quality and tidiness assessment with define, code, and test phases

Definition Phases:

1- Drop tweets with missing data in the expanded_urls.
2- Melt the 4 columns with the dog stages
3- Drop all rows containing retweets
4- Drop all rows containing replies
5- Drop all columns related to retweets
6- Drop all columns related to replies
7- Replace all lowercase words in the name column with the string "None"
8- Drop tweets with rating_denominator values that are NOT equal to 10.
9- Drop tweets that have rating_numerator >= 15.
10- Drop the rating_denominator column.

11- Rename the rating_numerator column to rating.
12- Replace the source string with the string between <a> and .
13- Create breed and confidence columns in the image predictions data frame and merge it with the archive data frame.
14- Merge the retweet_count, favorite_count, and user_count columns from the API data frame to the archive data frame, joining on tweet_id.
15- Change datatypes of timestamp to datetime, dog_stage to categorical, tweet_id to string, retweet_count to int, favorite_count to int, and user_count to int.

Store Cleaned Archive.

Stored cleaned archive data frame to a CSV file called twitter_archive_clean