Aigerim Shopenova
Data Analyst Nanodegree Program

Project: Wrangle and Analyze Data

In this report data gathering, data assessment and data cleaning processes are explained for the project.

Data Gathering

I gathered data from the following resources:

- Enhanced Twitter Archive named WeRateDogs, which was manually downloaded from Udacity's website
- The image predictions file for dog breeds, which was programmatically downloaded from Udacity's server
- JSON file with retweet count and favorite count was programmatically downloaded using the tweet IDs in the WeRateDogs Twitter archive. I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Data Assessment and Cleaning

Each data source, which was stored in a separated dataframe was assessed both visually and programmatically for quality and tidiness issues. Then, I cleaned data based on the observations and tested the data to check the code. Further, I combined the cleaned tables and stored in a master dataframe the obtained results.

Twitter archive

Because we only focused on original tweets, all rows containing non-null values in *retweeted_status_id, retweeted_status_user_id* and *retweeted_status_timestamp* were dropped. Then, it was found that *in_reply_to_status_id* and *in_reply_to_user_id* columns became empty, that is why I also dropped the columns.

Timestamp was converted from string format to datetime.

The 4 columns of dog stages for one variable does not conform to the rules of "tidy data". Therefore, I combined the columns into one called *dog_stages.*

It was found that there are cases with a rating numerator < 10 and >>10, which does not look like the truth. Therefore, I decided to drop the values with a rating numerator < 10 and >>10.

Also, it was found that there are 23 cases with the rating denominator != 10. I decided to remove the columns with the incorrect parameter.

The words in the name column, which do not look like names of dogs, were replaced by None.

There are 59 tweets with missing urls (expanded_urls - links to the tweet) in the enhanced twitter archive. I decided to drop the rows with missing data.

Image Predictions

The table itself was not cleaned. However, there are many tweets without a dog breed prediction, which was not removed. The dog breed prediction with the highest confidence level p1 was kept and combined with the archive table as the twitter table contains information about the dogs in the tweets. Also, p1 was renamed to explicitly explain the data and other predictions were dropped.

JSON file

The table itself was not cleaned too. The *retweet_count* and *favorite_count* columns were merged with the cleaned Twitter archive.

The merged dataframe containing data from the cleaned Twitter archive, Image Predictions file and JSON data on favorits and retweets were saved in 'master_twitter_archive_clean.csv'.