WE RATE DOGS TWITTER ARCHIVE
WRANGLE REPORT

Nosa Ogbeide-Ihama
Udacity Junior Data Analyst Nanodegree
Project 5

We Rate Dogs is a twitter account where followers dogs are rated. The wrangling project is to be carried out using this account. In carrying out this project, data required was gotten from three different sources which was then assessed, cleaned and analyzed. The aim of the project was to wrangle the datasets and derive insights from it.

Gathering Data

Three different datasets was used in this project:

1. Weratedogs twitter archive in a file named twitter_archive_enhanced.csv was downloaded manually provided by Udacity.
2. Tweet image predictions named image_predictions.tsv filewhich is hosted on Udacity's server was downloaded programmatically using request library.
3. The entire set of each tweet's json data in a file named tweet_json.txt gotten from querying twitter API using python's tweepy library. The favorite count and retweet count as well as tweet ID were then extracted from the file.

Assessing Data

These three pieces of data was assessed both manually and programmatically. The data was assessed manually by looking through each dataset to check for quality and/or tidiness issues while using codes like info(), describe(), duplicated(), unique()… to assess the data programmatically. A number of issues were found and noted down.

Cleaning Data

The quality and tidiness issues noted down in the assessment stage was being resolved in the cleaning stage. The twitter_archived_enhanced data had majority of the quality issues.

First, a copy was made of all three datasets, the tweet_id column of the datasets were converted to string. The timestamp column of twitter archive was converted to datetime. In_reply_to_status_id, In_reply_to_user_id, retweeted_status_id and retweeted_status_user_id had non-null values which were removed as original tweets were required, and then dropped.

Rating_denominator values were changed to 10. Values in all columns doggo, puppo, pupper and floffer that had none was changed to no_stage.

Missing values in expanded_urls column was dealt with and duplicated rows were dropped.

Incorrect names in name column were changed to ‘None’ as the names of the dogs could not be gotten from tweet text.

Rating_numerator column had inconsistent values and was changed manually. Correct ratings were collected from tweet texts. Ratings that could not be gotten from tweet text were dropped.

The four dog stage columns (doggo, puppo, pupper and floffer) were merged into one column called dog stage.

The JSON dataset containing the entire set of each tweet was merged with the twitter archived enhanced dataset.

Two extra columns were created for the image predictions table (dog breed and prediction confidence). These columns were gotten by conditionally selecting rows that were true for any of p1_dog, p2_dog or p3_dog columns. Where these three columns were true, p1 was chosen for dog breed and p1_conf for prediction confidence. Where all three columns were false, ‘none’ value was placed in dog breed column and 0 placed in prediction confidence table. Rows with none value in dog breed column was dropped.

These two columns (dog breed and prediction confidence) were then merged with twitter archive enhanced dataset. After merging, columns dog_breed and Prediction_confidence had null values which were then filled with ‘none’ and 0 respectively. The initial number of tweets collected were 2356 but after the cleaning process, 2083 tweets were left.