WeRateDogs Data Wrangling Project

WeRateDog is a Twitter handle that rate dog, upload dog image with a short description and give its followers a chance to like and re-tweet them. They are unique for using improper fractions as rating scores. We asked WeRateDog for their twitter-archive-enhanced and they mailed it to us named twitter-archive-enhanced. I programmatically downloaded, opened it, and observed it contained basic information but not everything. Twitter-archive-enhanced lacked essential information like favorite and re-tweet counts. Also, it doesn't contain information about the breed but thanks to Udacity for classifying the dog image uploaded using neural network procedure and saving it as image-prediction.

To create meaningful insight from this archive data, I needed more information such as re-tweet and favorite counts. So, I applied for a Twitter developer account which will allow me to query Twitter API with the help of the tweepy library. I fetched re-tweet and favorite counts for each tweet-id in the twitter-archive-enhanced and saved it as tweet-df.

Thereafter, I join all three data (image-prediction, twitter-archive-enhanced, and tweet-df) on tweet-id. I used inner join because not I could not get data for all the tweet-id in twitter-archive-enhanced.

Then, I did both visual and programmatic inspections and identify 8 quality and 2 tidiness issues.

Quality issues are:

1. Incomplete data (no favorite count, no re-tweet count)

2. Inconsistent way of representing missing values (NaN and None)

3. Some columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' contains little or no relevant observations

4. Timestamp datatype is object and tweet_id is 'int'

5. Erroneous rating score for most of the rating-numerator > 15 and the rating numerator == 0

6. Erroneous rating score for most of the rating-denominator != 10

7. Source written as a link instead of the exact name of the source

8. Wrong dog name ('a', 'such', 'quite' etc.)

Tidiness Issue

1. Dog age (doggo, fluffer, pepper, and puppo) is in four columns instead of one
2. Image prediction should be part of the twitter-archive-enhanced
3. Multiple breeds name (p1, p2, and p3) predicted for a dog

Thereafter, I wrote code to clean both the quality and tidiness issues. I then save the cleaned record as twitter-archive-master

Some of the insights I created are:

1. Are the top 10 most favored the top 10 most re-tweeted dog breed?
2. Is the dog breed with the highest re-tweet from the age class with the highest average rating?
3. Which tweet source has the highest average rating of dog?
4. What is the major source of the tweet?

Finally, I downloaded the picture of the top five dog breeds with high favorite counts and used it to write a blog post.