WRANGLE REPORT

by Ikechukwu Ogbuchi

We had 3 datasets for this project which were all loaded onto Jupyter notebook using the pandas read function and I looked at these datasets for quality, naming of columns and checking out for missing values. I detected 10 quality issues with the dataset from both visual and programmatic assessments and proceeded to correct these datasets for analysis.

I discovered the following issues:

1. Column values not well defined in image predictions p1, p2
2. Columns with many missing values like "retweeted_status_id", "retweeted_status_timestamp", "in_reply_to_status_id", "contributors","retweeted_status_user_id", "coordinates" and others.
3. Presence of 'None' for missing records, instead of pandas 'NaN'
4. Invalid values for 'name' column like 'an', 'a', 'such', 'Bo',
5. 'tweet_id' and 'id' columns are of type int rather than str
6. The 'text' column contains both original tweets and retweets, while we were required to work with just original tweets
7. Invalid data type for timestamp column.
8. Some predictions are false on dogs

And these tidiness issues:

1. Multiple dog stages column present
2. Datasets are separate and need to be merged into one after all cleaning

This project is interesting as it is about dogs and tweets from dog lovers. Here is a picture of a lovely golden retriever:

Fig 1. Golden Retriever

I resolved the identified issues by making copies of each dataset using the copy function and then began to remove irrelevant columns using the python drop function, and used the rename function rename columns that I felt were not properly defined which included the p1,p2 and p3 column names.

I also removed all of the columns in dog predictions that were not dogs or false. I had used where function to compare predictions from each prediction column then filtered the entire dataset with true values.

Replaced all None values with NaN for easier manipulation, removed rows with invalid names like a, the, such and ensured I removed all tweets with RT. Further converted the timestamp format in data from object to datetime.

After making the required changes, I renamed the id column in the twitter data to “tweet_id” and merged the three datasets on tweet_id which was the common column to all three variables.

To ensure I would not lose any changes, I kept backing up my code by downloading a copy and checking that my Kernel was responding. I checked the timestamp as well and sorted it to ensure that it did not include data after Aug 1, 2017. After that I began visualization and insights.