Wrangle Report

By Ashish

The project 'WeRateDogs' was very challenging, and I learned a about the data cleaning process and the Twitter API.

This report briefly describes my wrangling efforts,

Gathering Data:

- **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

- **The tweet image predictions**, i.e., what breed of dogs are present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

- **Twitter API & JSON:** by 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. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing Data:

After gathering all the data in three data frames. I started visual assessment by printing all the three data frames. In visual assessment many spelling mistakes, data type issue was found. In programmatic assessment *info, value_counts, duplicated* was used to identify quality and tidiness issues.

Cleaning Data:

Now the cleaning part was most challenging part in this project. It was divided into three steps Define, Code, Test.

I made a copy of all the three data frames and started cleaning process. I then converted columns to a proper data format, primarily changing the timestamp data into datetime objects, tweet_id from a number into a string and the rating columns into float objects. I also addressed quality issues in the Prediction columns of the Image Prediction dataframe. Utilizing the pandas library str.replace() and str.title() functions, I removed the underscore between the words and

capitalized the letter in each word to make a more cohesive table. Then at last pd.merge() function was used to merge all three cleaned data frames

The most challenging part was to get the dog ratings. Took help from internet, slack forum. Then it was done through lambda function.

Overall it was a great learning for me in Data Science field.