Ehsan Jafari-Shirazi

Wrangling Data Project –
Wrangling Twitter Archived Dataset: WeRateDogs

During this data wrangling project, three different data sources were gathered, assessed the quality and tidiness of the data, cleaned by using Python and its libraries within Jupyter Notebook platform, and finally the cleaned version of the data sources were merged by using an inner join to have the dataset ready for the next portion of the project which was the data analysis and visualization of the findings/insights.

The first data source, twitter_archive_enhanced, was given by Udacity in the form of a csv file. This dataset contained basic tweet data such as tweet ID, timestamp, text, etc. for all 5000+ of tweets as archived in the WeRateDogs Twitter account.

The second source of data was an image prediction file where the dogs' breeds were predicted by running each tweet image through a neural network that can classify breeds of dogs. This dataset was programmatically downloaded as a tsv file using the Requests Python library.

The last piece of data was tweet's JSON data which was the most challenging part of the gathering step. To gather this dataset, the tweet IDs from the twitter_archive_enhanced dataset was used to query the Twitter API for each tweet's JSON data by using the Python's Tweepy library. This main purpose of gathering this dataset was to obtain each tweet's retweet and favorite counts.

After gathering all the three sources of data, the data assessment step was started. For this step, I visually and programmatically assessed all the three datasets with regards to data quality and tidiness issues. A summary of the data quality and tidiness issues identified are as following:

Quality Issues:

- ✓ Retweeted rows that need to be removed.
- ✓ Tweets that have no images
- ✓ Inaccuracy, mislabeling or missing dog names in the 'name' field
- ✓ Missing values in dog stages (where showing as 'None')
- ✓ Inaccurate Rating data: When 'Text' field includes more than one #/#, sometimes the first occurrence erroneously have used for the rating numerators
- ✓ Rating numerators which have decimals sometimes not showing full float
- ✓ 'Text' column containing extra characters after '&'
- ✓ Erroneous datatypes for the following columns:
    - tweet_id
    - timestamp
    - dog stages
    - in_reply_to_status_id
    - in_reply_to_user_id

Tidiness Issues:

- ✓ Dog "stage" is a variable and should be in one column rather than four columns: doggo, floofer, pupper, puppo
- ✓ Parse the datetime information from one column (timestamp) into separate columns: Date and Time. This is to follow the following Tidiness rule: Avoid multiple variables to be stored in one column.
- ✓ Join the three dataset to create a new dataset which is the final combined data.
- ✓ Keep and store the clean version of the three dataframes: to follow the following Tidiness rule: "Each type of observational unit forms a table."

After this, the final step of data wrangling project was to clean the data by addressing each of the quality and tidiness issues identified as above in the assessment step.

Lastly, the cleaned datasets were combined by using an inner join to create a new dataset which was called twitter_archive_master. It should be noted that the cleaned version of the three original datasets were stored in a folder called “Stored_Clean_Data” to keep tidiness of the data, as multiple tables were required.