Data Wrangling Report

WeRateDogs is a twitter handle that posts about pictures and tweets about dogs along with a rating out of ten. This handle currently (June 2020) has around 8.7M followers and has a global base of followers.

For this project, datasets are gathered from three different sources. The first of them is a dataset from twitter archives which can be loaded from a CSV file, the second one is the image-prediction dataset which is a TSV file which is loaded using the request library and the following URL:
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
The last dataset is collected from each tweet's JSON file using Python's Tweepy library. This is loaded from the tweet_json.txt file.

For each dataset, the first phase is gathering and loading data. The obtained dataset is loaded into a dataframe. For this project, the tweet_id is used as an index to uniquely identify tweets as well as to help with merging the datasets in the later stages.

In the second phase, i.e. Data Assessing, now the datasets are analyzed in order to list any quality or tidiness issues in the raw data. These inconsistencies could affect further insights, so it is necessary to remove them well in advance. The analysis could be done using various functions like info(), describe(), isnull(), etc. as well as by plotting a box /scatter plot to locate outliers or random values.

For each dataset, the following are some data quality and tidiness issues.

Dataset 1: Twitter Archives

Data Quality Issues

1) Some of the tweets in the dataset are retweets, so these tweets must be dropped.

2) Some of the tweets in the dataset are replies, so these tweets must be dropped.

3) Timestamp column has object datatype instead of the correct Date-Time format.

4) Some tuples in the dataset contain multiple, composite URLs instead of a single URL in the expanded_urls column.

5) Dropping values in the name column like 'a', 'the', 'such', etc. which are irrelevant and all have lowercase characters.

6) Fix incorrect numerator and denominator ratings as well as normalize all ratings to get a standard reference.

Tidiness Issues

1) To drop unnecessary columns from the dataset, as they don't contribute enough in finding out any insights. For this dataset, these columns are dropped: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and source.

2) Instead of having separate columns for dog stages like doggo/ floofer/ pupper/ puppo, it could be represented in a single column.

Dataset 2: Image Predictions

Data Quality Issues

1) The dog breed prediction columns (p1,p2,p3) contain a '_' between every two words. So, it needs to be stripped off from the column values.

2) To drop duplicate image URLs (66) from the dataset.

Tidiness Issues

1) Most of the columns in the dataset have confusing column names. So they need to be renamed such that most people would understand them.

Dataset 3: Tweets JSON file

After assessing this dataset, no major discrepancies weren't found. So, data cleaning is not needed for this dataset.

The third and the last phase, Data cleaning, deals with solving the issues found in the last phase to make data useful for further analysis. This involves two steps:

1. Code- To write code to remove the specified issue.
2. Test- To check whether the issue still persists even after the code step.

Once all these three phases are completed, now we merge the datasets using the reduce method. The merging takes place on both the ends based on the tweet_id and this final dataframe is loaded into a CSV file using the to_csv method of pandas library. The final dataset contains 1809 rows and 21 columns.